Quantification Learning with Applications to Mortality Surveillance

by

Jacob Fiksel

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

April, 2020

Abstract

This thesis is motivated by estimating the cause specific mortality fraction (CSMF) for children deaths in Mozambique. In countries where many deaths are not assigned a cause of death, CSMF estimation is often performed by performing a verbal autopsy (VA) for a large number of deaths. A cause for each VA is then assigned via one or more computer coded verbal autopsy (CCVA) algorithms, and these cause assignments are aggregated to estimate the CSMF. We show that CSMF estimation from CCVAs is poor if there is substantial misclassification due to CCVAs being informed by non-local data. We develop a parsimonious Bayesian hierarchical model that uses a small set of labeled data that includes deaths with both a VA and a gold-standard cause of death. The labeled data is used to learn the misclassification rates from one or multiple CCVAs, and in-turn these estimated rates are used to produce a calibrated CSMF estimate. A shrinkage prior ensures that the CSMF estimate from our Bayesian model coincides with that from a CCVA in the case of no labeled data. To handle probabilistic CCVA predictions and labels, we develop an estimating equations approach that uses the Kullback-Liebler loss-function for transformation-free regression with a compositional outcome and predictor. We then use Bayesian updating of this loss function, which allows

for calibrated CSMF estimation from probabilistic predictions and labels. This method is not limited to CSMF estimation and can be used for general quantification learning, which is prevalence estimation for a test population using predictions from a classifier derived from training data. Finally, we obtain CSMF estimates for child deaths in Mozambique by applying all of the developed methods to VA data collected from the Countrywide Mortality Surveillance for Action (COMSA)-Mozambique and VA and gold-standard COD data collected from the Child Health and Mortality Prevention project.

Thesis Committee

Primary Readers

Abhirup Datta (Primary Advisor)
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Robert Scharpf
Associate Professor
Department of Oncology
Johns Hopkins University School of Medicine

Agbessi Amouzou
Associate Professor
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Scott L. Zeger
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Tom Louis
Professor Emeritus
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Jennifer A. Deal
Assistant Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

I'd first like to say thank you to my advisor Abhi. Besides truly being one of the smartest and hardest working people I've met, you're just incredibly kind and fun to be with. You always made it a point to make sure that I took time off my thesis work to spend valuable time with my friends and family, and I felt that every decision you made as my advisor was with my best interest in mind. You helped me to become a better statistician, writer, and communicator, and trusted me with important analyses and presentations that were instrumental in my development as a PhD student. This thesis wouldn't have been possible without your guidance.

To Rob Scharpf, thank you for taking me in from day 1 of the PhD. While our work together did not end up in this thesis, I have learned an incredible amount from you as a mentor. You always supported me in all my decisions, even when those decisions involved doing research outside of the lab. You have also served as a guide on how to balance family and academic life, and I can only hope to be as dedicated to my (future) family as you are to yours. Above all, I have enjoyed our friendship over the past 5 years, and I hope this continues into the future.

To Scott Zeger, thank you for being a mentor to both Abhi and I throughout

this thesis work. Your suggestions were invaluable, as well as your ability to champion our approach during the meetings with the COMSA and CHAMPS teams. I also want to thank you for always having an open home, and inviting me over for several lovely dinners with your family throughout the PhD.

Thank you to Agbessi. You have been a wonderful collaborator in working with us to develop and integrate these methods into COMSA's statistical workflow. Thank you for ensuring that we had all of the data we needed, and giving me the opportunity to travel to Mozambique to present this work.

Thank you to Jennifer and Tom, for being alternate committee members, and also great supporters for this work. I also want to say thank you to Kala Visvanathan for being part of my oral exam committee.

To Victor–I also have to say thank you for welcoming me into the lab from day 1. I really appreciated getting to be a part of your groundbreaking studies on using cell-free DNA for early cancer detection, and getting to work in a lab that so clearly values both wet-lab and dry-lab work is one of the reasons why Hopkins is such a special place. The weekly lab lunches weren't so bad either. Although my thesis work did not come from the lab, I cannot say how much I appreciate you always treating me as part of the Velculescu Lab family.

Thanks to Sara Alcorn, with whom I got to collaborate on important work for improving palliative radiation treatments. It was really great working with you, and also without you, I would never have ended up working with Scott, who was the one who introduced me to Abhi.

I'd like to thank my funding source, the EBA program, as well as my EBA advisors Dani and Ravi, and EBA administrators Brian and Monique. I really

enjoyed attending the research-in-progress meetings over the past 4 years, as well as getting your feedback on my work. The program has opened my eyes to incredibly important scientific and public health research on aging and has transformed my thinking as a public health statistician.

Thank you to everyone on the COMSA and CHAMPS team, especially to Emily Wilson who has to clean all of the verbal autopsy data, and has done an incredible job in getting data to us on short notice.

Thanks to Karen for her leadership as department chair, especially during the covid-19 pandemic. Thanks to Brian and Hongkai, the two graduate program directors during my time as a PhD student. Also thanks to Margaret, Leah, and Marie for allowing me to build my teaching skills while serving as their TA.

Thanks to all the department staff, especially Mary Joy, who everyone knows is the glue that holds this department together. Also thank you to all of staff who have built and provide support for the cluster, this PhD really wouldn't have been possible without using the cluster.

Of course, thanks to Patty Hubbard, who always stopped by my office to chat, and made seminar special by making sure we had the best pastries.

Thank you to my undergraduate advisor, Jo Hardin. I wouldn't have gone into this field without you, and I'm incredibly glad we've remained in contact and even got to write a paper together. I'm looking forward to now being your colleague!

Thanks to my office mates in E3035, there's no one else I'd rather procrastinate with while drinking our freshly ground coffee.

Thanks to my boys, Ben, Matt, and Lamar for sticking together and being such good friends over the past 5 years–I really don't know if I would have made it through the first year without you.

Thank you to all the great dogs in my life over the past five years, especially Jonas, Hilde, and Teddy. And thank you to Dan and Pat for giving me an incredible apartment to live in over the past four years.

Thanks to Emily for being the absolute best girlfriend. I really appreciate all of your support and love, especially during the last incredibly stressful month of writing this thesis. I can't wait for more food and outdoors adventures with you.

Finally, thank you to my family, especially my mom, my dad, and my step-mom. Thanks for all of your support and love not only during the PhD, but always. I love you lots and can't wait to celebrate together.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

High-quality cause-of-death (COD) information is crucial for governments and policy makers to evaluate progress towards development goals and to guide new policies (Lopez and Setel, 2015). However, this high-quality information is lacking for 65% of the world's population (Nichols et al., 2018), due to the fact that few complete diagnostic autopsies are performed in low-and middle-income countries (LMICs). As a practical method of obtaining COD information, LMICs have been adapting the use of verbal autopsies (VAs) (Fottrell and Byass, 2010). A VA, which involves a detailed interview with a close relative or neighbor of the deceased, will result in information about a list of hundreds of symptoms that the deceased may or may not have been experiencing before their death.

Demographic surveillance systems, like the Countrywide Mortality Surveillance for Action (COMSA) – Mozambique, necessitate collecting COD information for hundreds to thousands of deaths. Having two physicians review each VA to assign a COD, as is standard practice (Soleman, Chandramohan,

and Shibuya, 2006), will be too time consuming and costly. Instead, to produce COD assignments from VA data, surveillance systems are turning to automated, computer-coded classifiers for VA algorithms (CCVA) such as InSilicoVA (McCormick et al., 2016), InterVA-4 (Byass et al., 2012), the Naive Bayes Classifier (NBC) for Verbal Autopsies (Miasnikof et al., 2015), and the expert algorithm for verbal autopsy (EAVA) (Kalter, Perin, and Black, 2016). COD assignments from a CCVA are aggregated to obtain a cause specific mortality fraction (CSMF) estimate.

Chapter 2 of this thesis demonstrates that simply using the aggregated COD assignments from one or multiple CCVAs may produce inaccurate CSMF estimates, due to the fact that CCVAs are imperfect classifiers. The reason for this is that the local context of a VA, such as how respondents describe a symptom or disease, may differ between countries. Thus, a CCVA trained using VA data from India would be expected to perform poorly when applied to VA data from Mozambique.

We develop a Bayesian framework that allows using information on the inaccuracy of the classifier from a small validation set from the target population. The validation set is often ongoing collection of a small set of hospital deaths where both a VA and a GS-COD. This, alongside the larger set of nationally representative community deaths where just a VA is available are jointly used in a hierarchical model. The hospital deaths are used to learn the sensitivities and specificities, which we refer to as misclassification rates, for CCVAs. The misclassification rates are then used to calibrate the aggregated CCVA CSMF estimates. Importantly, our framework uses a shrinkage prior that guarantees

the calibrated CSMF estimate will coincide with the uncalibrated CSMF estimate (or average of the uncalibrated CSMF estimates from multiple CCVAs) when there is little-to-no hospital death data. We also develop an ensemble approach, which combines information from multiple CCVAs to estimate a single CSMF. This ensemble estimate prevents us from having to decide on which CCVA to use for calibration, and performs at par with the most accurate classifier. We develop a simple and fast Gibbs sampler for obtaining posterior samples, which is implemented in the 'CalibratedVA' R-package.

However, there are two short comings of the method presented in Chapter 2. First, the method only allows for a single-cause assignment from a CCVA (single-cause-VA). Algorithms such as InSilicoVA give probabilistic individual cause assignments, and simply using the most-likely COD for each individual as the cause assignment results in a loss of information. Second, the method requires that each hospital death has a *known* GS-COD (single-cause-GS-COD). If this GS-COD is determined by an expert panel, as it is for minimally invasive autopsies, there may be uncertainty in the final cause assignment. Thus, we would like to extend the method to handle a probabilistic GS-COD (multi-cause-GS-COD).

Chapter 3 develops the statistical tools needed to shift from the single-cause-VA-single-cause-GS-COD calibration to the multi-cause-VA-multi-cause-GS-COD calibration. Recognizing that the multi-cause-VA and multi-cause-GS-COD for each death are *compositional* vectors, we develop a model for transformation-free linear regression for compositional outcomes and predictors, which we call the direct regression model. While we apply this model to

multi-cause-VA-multi-cause-GS-COD in Chapters 4 and 5, the direct regression model is not limited to this particular application. We apply the direct regression model to compositional data collected from both education and medical research. The direct regression model is simple to interpret, which is not the case for the compositional regression models developed by Chen, Zhang, and Li (2017) and Alenazi (2019). Our method also seamlessly allows for 0s and 1s in the compositional data. Interestingly, when both the compositional outcome and compositional predictor are categorical and thus only have 0s and 1s, as in the single-cause-VA-single-cause-GS-COD scenario, our model reduces to a risk-prediction model for a multinomial outcome.

Chapter 4 frames the problem of CSMF estimation from aggregation of CCVA COD assignments as a *quantification* (Forman, 2005) problem. Quantification is the task of predicting the population distribution (prevalence) of unobserved true outcomes (labels) based on observed covariates (Forman, 2005; González et al., 2017). As with CSMF estimation from a CCVA, quantification is often performed by predicting individual outcomes (CODs) using covariates (VA symptoms), and then aggregating these predicted outcomes. However, simply aggregating predicted labels from a classifier to estimate the population prevalence ignores the fact that classifiers are often imperfect.

We develop the Generalized Bayesian Quantification Learning (GBQL) method that allows for quantification from both single-class and multi-class (probabilistic) classifier output. Our method is based on the estimating equations approach in Chapter 3 that uses the Kullback-Liebler loss-function. We jointly estimate the population label prevalence and classifier misclassification

rates by incorporating loss functions for both labeleled and unlabeled data. Data are allowed to have compositional (probabilistic) labels, which allows for quantification using multi-cause-GS-COD data. The important work developed by Bissiri, Holmes, and Walker (2016) allows for Bayesian updating of posteriors using loss-functions. We can thus incorporate the shrinkage prior developed in Chapter 2 into this model, while also modeling the probabilistic classifications and labels. In addition, we demonstrate how different choices of shrinkage priors ensures that, in the absence of labeled test data, quantification from our method shrinks to different existing quantification methods like classify & count (CC) (Forman, 2005) or probabalistic average (PA) (Bella et al., 2010). As in Chapter 2, we develop an ensemble method that produces a single prevalence estimate, but classifier-specific misclassification rate estimates. This ensemble method is based on minimizing the average loss across all of the classifiers.

Chapter 5 concludes the thesis by applying the methods from the previous chapters to estimate the CSMF for child (1-59 months old) deaths in Mozambique. We use multi-cause-VA output from InSilicoVA and EAVA for 989 VAs collected by COMSA. To obtain a set of labeled data, we use data from the Child Health and Mortality Prevention (CHAMPS) project. CHAMPS is an ongoing surveillance project that performs a minimally invasive autopsy (MIA), also known as a minimally invasive tissue sample (MITS) (Byass, 2016) to determine the COD with high precision. A VA for each death that occurs within a CHAMPS site (*CHAMPS Cause of Death Data*) is also conducted. MITS COD assignments have been shown to be very accurate compared to complete

diagnostic autopsies (Castillo et al., 2016). However, because MITS COD assignments are decided on by an expert human panel, there may be some uncertainty in the final cause assignment.

The direct regression model allows us to use the MITS and VA data collected on child deaths that occurred at the CHAMPS sites (including Mozambique) to estimate the uncalibrated misclassification rates of InsilicoVA and EAVA. These estimates reveal substantial classification errors for both algorithms cautioning against the use of the raw CSMF estimates as they are likely to be very biased. We use the misclassification matrices to produce calibrated VA CSMF estimates for child deaths in Mozambique. We use the GBQL framework to handle uncertainty in MITS COD classification, as well as to incorporate probabilistic individual COD predictions from VA algorithms. We demonstrate a complete workflow of the methodology that first estimates the raw CSMF estimates and misclassification rates, combines them to produce a single calibrated ensemble CSMF estimate, and provides quantitative model comparison metrics to compare and choose between the raw and calibrated CSMF estimate.

References

Lopez, Alan D and Philip W Setel (2015). "Better health intelligence: a new era for civil registration and vital statistics?" In: *BMC medicine* 13.1, p. 73.

Nichols, Erin K, Peter Byass, Daniel Chandramohan, Samuel J Clark, Abraham D Flaxman, Robert Jakob, Jordana Leitao, Nicolas Maire, Chalapati Rao, Ian Riley, et al. (2018). "The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0". In: *PLoS medicine* 15.1.

Fottrell, Edward and Peter Byass (2010). "Verbal autopsy: methods in transition". In: *Epidemiologic reviews* 32.1, pp. 38–55.

Soleman, Nadia, Daniel Chandramohan, and Kenji Shibuya (2006). "Verbal autopsy: current practices and challenges". In: *Bulletin of the World Health Organization* 84, pp. 239–245.

McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark (2016). "Probabilistic cause-of-death assignment using verbal autopsies". In: *Journal of the American Statistical Association* 111.515, pp. 1036–1049.

Byass, Peter, Daniel Chandramohan, Samuel J Clark, Lucia D'ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. (2012). "Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool". In: *Global health action* 5.1, p. 19281.

Miasnikof, Pierre, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha (2015). "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths". In: *BMC medicine* 13.1, p. 286.

Kalter, Henry D, Jamie Perin, and Robert E Black (2016). "Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death". In: *Journal of global health* 6.1.

Chen, Jiajia, Xiaoqin Zhang, and Shengjia Li (2017). "Multiple linear regression with compositional response and covariates". In: *Journal of Applied Statistics* 44.12, pp. 2270–2285.

Alenazi, Abdulaziz (2019). "Regression for Compositional Data With Compositional Data as Predictor Variables With or Without Zero Values". In: *Journal of Data Science* 17.1, pp. 219–237.

Forman, George (2005). "Counting positives accurately despite inaccurate classification". In: *European Conference on Machine Learning*. Springer, pp. 564–575.

González, Pablo, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz (2017). "A review on quantification learning". In: *ACM Computing Surveys (CSUR)* 50.5, p. 74.

Bissiri, Pier Giovanni, Chris C Holmes, and Stephen G Walker (2016). "A general framework for updating belief distributions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 1103–1130.

Bella, Antonio, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana (2010). "Quantification via probability estimators". In: *2010 IEEE International Conference on Data Mining*. IEEE, pp. 737–742.

Byass, Peter (2016). "Minimally invasive autopsy: a new paradigm for understanding global health?" In: *PLoS medicine* 13.11.

CHAMPS Cause of Death Data. https://champshealth.org/cause-of-death-data-visualization/.

Castillo, Paola, Miguel J Martínez, Esperança Ussene, Dercio Jordao, Lucilia Lovane, Mamudo R Ismail, Carla Carrilho, Cesaltina Lorenzoni, Fabiola Fernandes, Rosa Bene, et al. (2016). "Validity of a minimally invasive autopsy for cause of death determination in adults in Mozambique: an observational study". In: *PLoS medicine* 13.11.

Chapter 2

Regularized Bayesian transfer learning for population-level etiological distributions

2.1 Introduction

Verbal autopsy – a survey of the household members of a deceased individual, act as a surrogate for medical autopsy report in many countries. Computer-coded verbal autopsy (CCVA) algorithms are high-dimensional classifiers that predict cause of death from these high-dimensional family questionnaires which are then aggregated to generate national and regional estimates of cause-specific mortality fractions (CSMF). These estimates may be inaccurate as CCVA are usually trained using non-local information not representative of the local population of interest. This problem is a special case of *transfer learning*, a burgeoning area in statistics and machine learning.

Classifiers trained on *source domain* data tend to predict inaccurately in a *target domain* different from the source domain in terms of marginal and

conditional distributions of the *features* (covariates) and *labels* (responses) (Shimodaira, 2000). Various *domain adaptation* strategies have been explored for transfer learning of generic classifiers which adjust for this distributional differences between the two domains. We refer the readers to Weiss, Khoshgoftaar, and Wang, 2016 and Pan and Yang, 2010 for a comprehensive review of transfer learning for classification problems. We focus on the setting where there is abundant labeled source domain data, abundant unlabeled target domain data, and limited labeled target data. Transfer learning approaches pertaining to this setting include multi-source domain adaptation (CP-MDA, Chattopadhyay et al., 2012), neural networks (TCNN, Oquab et al., 2014), adaptive boosting (TrAdaBoost, Dai et al., 2007; Yao and Doretto, 2010), feature augmentation method (FAM, Daumé III, 2009), spectral feature alignment (SFA, Pan et al., 2010) among others.

All of the aforementioned transfer learning classification approaches are motivated by applications in image, video or document classification, text sentiment identification, and natural language processing where individual classification is the goal. Hence, they usually focus on the individual's (e.g. a person's or an image's) classification within a target domain (e.g. a particular population) with training performed in data from a different source domain.

Social and health scientists such as epidemiologists are often more interested with understanding etiological distributions at the population-level rather than classifying individuals. For example, we aim to estimate national and regional estimates of cause-specific fractions of child mortality. Hence,

our goal is not individual prediction but rather transfer learning of population-level class probabilities in the target domain. None of the current transfer learning approaches are designed to directly estimate population-level class membership probabilities.

Additionally, the extant transfer learning approaches rely on large source domain databases of millions of observations for training the richly-parameterized algorithms. The sample sizes of datasets in epidemiology are typically orders of magnitude smaller. Most epidemiological applications use field data from surveys, leading to databases with much smaller sample sizes and yet with high-dimensional co sets (survey records). For example, in our application, the covariate space is high-dimensional ($\sim 200 - 350$ covariates), the 'abundant' source domain data has around ~ 2000 samples, while the local labeled data can have as few as $\sim 20 - 100$ samples. Clearly, in such cases, the local labeled data is too small to train a classifier on a high-dimensional set of covariates, as the resulting estimates will be highly variable. A baseline classifier trained on the larger source domain data will tend to produce more stable estimates, but the high precision will come at the cost of sacrificing accuracy if the source and target domains differ substantially.

Our parsimonious solution to this bias-variance trade-off problem is to use the baseline classifier trained on source-domain information to obtain an initial prediction of target-domain class probabilities, but then refine it with the labeled target-domain data. We proffer a hierarchical Bayesian framework that unifies these two steps. With C classes and S-dimensional covariates, the advantage of this new approach is that the small labeled data for the target

domain is only used to estimate the $C \times C$ *confusion matrix* of the *transfer error* (misclassification) rates instead of trying to estimate $\mathcal{O}(SC)$ parameters of the classifier directly from the target-domain data. Since $S \gg C$, this approach considerably reduces the dimensionality of the problem. To ensure a stable estimation of the confusion matrix, we additionally use a regularization prior that shrinks the matrix towards identity unless there is substantial transfer error. We show that, in the absence of any target domain labeled data or in case of zero transfer error, posterior means of class probability estimates from our approach coincide with those from the baseline learner, establishing that the naive estimation that ignores transfer error is a special case of our algorithm. We devise a novel, fast Gibbs sampler with augmented data for our Bayesian hierarchical model.

We then extend our approach to one that uses an ensemble of input predictions from multiple classifiers. The ensemble model accomplishes method-averaging over different classifiers to reduce the risk of using one method that is inferior to others in a particular study. We establish a theoretical result that the class probability estimates from the ensemble model coincides with that from a classifier with zero transfer error. A Gibbs sampler for the ensemble model is also developed, as well as a computationally lighter version of the model that is much faster and involves fewer parameters. Simulation and data analyses demonstrate how the ensemble sampler consistently produces estimates similar to those produced by using our transfer learning on the single best classifier.

Our approach is also post-hoc, i.e., only uses pre-trained baseline classifier(s), instead of attempting to retrain the classifier(s) multiple times with different versions of training data. This enables us to use publicly available implementations of these classifier(s) and circumvents iterative training runs of the baseline classifier(s) which can be time-consuming and inconvenient in epidemiological settings where data collection continues for many years, and the class probabilities needs to be updated continually with the addition of every new survey record. The post-hoc approach also ensures we can work with non-statistical classifiers that do not use a training data but some sort of source domain information (e.g. CCVA algorithms InterVA and EAVA).

The rest of the manuscript is organized as follows. We present the motivating application in Section 2.1.1. In Sections 2.2 and 2.3, we present the methodology and its extension to the ensemble case. Section 2.9.1 presents an EM algorithm approach to obtain maximum a posteriori (MAP) estimates for the model, as a fast alternative to the fully Bayesian approach adopted earlier. Section 2.4 considers the extension where class probabilities can be modeled as a function of covariates like age and sex, and spatial regions. Section 2.5 presents simulation results. Section 2.6 returns to the motivating dataset and uses our transfer learning model to estimate national CSMFs for children deaths in India and Tanzania. We end the manuscript in Section 2.7 with a discussion of limitations and future research opportunities.

2.1.1 Motivating dataset:

In low and middle income countries, it is infeasible to conduct full autopsies for the majority of deaths due to economic and infrastructural constraints, and/or religious or cultural prohibitions against autopsies (AbouZahr et al., 2015; Allotey et al., 2015). An alternative method to infer the cause (or "etiology") of death (COD) is to conduct *verbal autopsy* (VA) – a systematic interview of the relatives of the deceased individual – to obtain information about symptoms observed prior to death (Soleman, Chandramohan, and Shibuya, 2006). Statisticians have developed several specialized classifiers that predict COD using the high-dimensional VA records as input. Examples include Tariff (James, Flaxman, and Murray, 2011; Serina et al., 2015), InterVA (Byass et al., 2012), InSilicoVA (McCormick et al., 2016), the King and Lu method (King, Lu, et al., 2008), EAVA or expert algorithm (Kalter et al., 2015), etc. Software for many of these algorithms are publicly available, e.g., Tariff (Li, McCormick, and Clark, 2018c), InSilicoVA (Li, McCormick, and Clark, 2018a), InterVA (Thomas et al., 2018) and the openVA R-package (Li, McCormick, and Clark, 2018b) has consolidated most of these individual software into a single package. Generic classifiers like random forests (Breiman, 2001), naive Bayes classifiers (Minsky, 1961) and support vector machines (Cortes and Vapnik, 1995) have also been used (Flaxman et al., 2011; Miasnikof et al., 2015; Koopman et al., 2015) for classifying verbal autopsies. Predicted COD labels for each VA record in a nationally representative VA database is aggregated to obtain national cause specific mortality fractions (CSMF) – the population-level class membership probabilities, that are often the main

quantities of interest for epidemiologists, local governments, and global health organizations.

Formally, a CCVA algorithm is simply a classifier using the $S \times 1$ covariate vector (VA report) $\mathbf{s}$ to predict c – one of C possible COD categories. Owing to the high-dimensionality of the covariate space (VA record consists of responses to $200 - 350$ questions), learning this mapping $P(c \mid \mathbf{s})$ requires substantial amount of *gold standard* (labeled) training data. Usually in the country of interest, VA records are available for a representative subset of the entire population, but gold standard cause of death (GS-COD) is ascertained for only a very small fraction of these deaths. In other words, there is abundant unlabeled data but extremely limited labeled data in the target domain. The ongoing project Countrywide Mortality Surveillance for Action (COMSA) Mozambique typify this circumstance, where, in addition to conducting a nationally representative VA survey, researchers will have access to gold standard COD for a small number of deaths from one or two local hospitals using minimally invasive autopsies (MIA) (Byass, 2016). Budgetary constraints and socio-cultural factors unfortunately imply that only a handful of deaths can eventually be autopsied (up to a few hundred).

Lack of sufficient labeled target-domain data implies that CCVA classifiers need to be trained on non-local data like the publicly available Population Health Metrics Research Consortium (PHMRC) Gold Standard VA database (Murray et al., 2011b), that has more than $10,000$ paired physician and VA assessments of cause of death across 4 countries. However, there exists considerable skepticism about the utility of CCVA trained on non-local data as

cause-symptom dynamics are often local in nature (McCormick et al., 2016; Flaxman et al., 2018). To illustrate the issue, in Figure ??, we plot the confusion matrices between the true COD of the PHMRC child cases in Tanzania against the predicted COD for these cases using two CCVA algorithms, Tariff and InSilicoVA, both trained on all PHMRC child data non-local to Tanzania. Both matrices reveal very large transfer errors, some as high as 60%. The

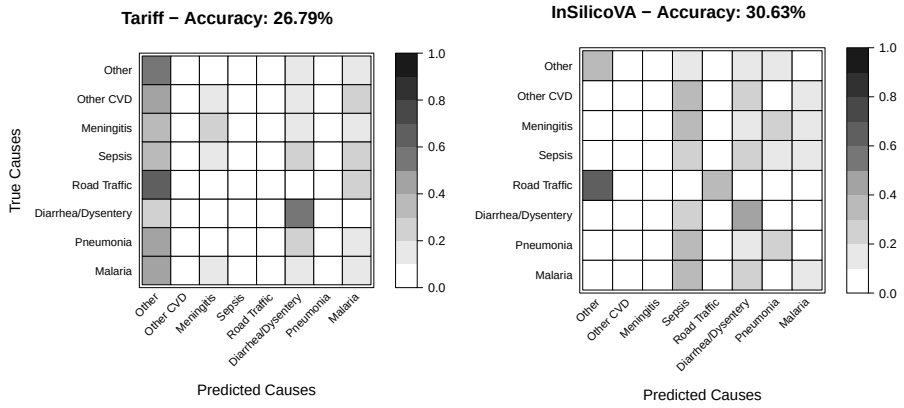


Figure 2.1: Confusion matrices for PHMRC child cases in Tanzania using Tariff and InSilicoVA trained on all cases outside of Tanzania. CVD is abbreviation for cardio-vascular diseases.

large transfer errors indicate that the naive estimates of population-level class probabilities from CCVA classifiers trained on non-local source data are likely to be inaccurate thereby highlighting the need for transfer learning in this application. Additionally, like for any other application area, there exists considerable disagreement about which CCVA algorithm is the most accurate (Leitao et al., 2014; McCormick et al., 2016; Flaxman et al., 2018). In our experience, no method is universally superior, and a robust ensemble transfer learning approach guarding against use of inaccurate classifiers is desirable.

Table 2.1: Glossary of acronyms used in the manuscript

Acronym	Full form	Acronym	Full form
VA	Verbal autopsy	PHMRC	Population Health Metrics Research Consortium
CCVA	Computer coded VA	COMSA	Countrywide Mortality Surveillance for Action
COD	Cause of Death	CSMF	Cause Specific Mortality Fraction
CSMFA	CSMF accuracy	GS-COD	Gold-standard Cause of Death

2.2 Transfer learning for population-level class probabilities

2.2.1 Naive approach

Let $\mathbf{p} = (p_1, p_2, \ldots, p_C)'$ denote the true population-level class probabilities in a target domain $\mathcal{D}_T$ where we have abundant unlabeled covariate data, which we denote by $\mathcal{U}$, and a very small labeled data $\mathcal{L}$ of paired labels and covariates. For a covariate vector $\mathbf{s}$, we can write $p_i = P(G(\mathbf{s}) = i)$ where G denote the true (gold standard) class-membership. Also, let $A(\mathbf{s}) = A(\mathbf{s} \mid \mathcal{G})$ denote the predicted class membership from the baseline classification algorithm A trained on some large labeled dataset $\mathcal{G}$ in a source domain $\mathcal{D}_S$ different from $\mathcal{D}_T$. If we do not use any transfer learning, the naive estimate of $\mathbf{p}$ from A is given by

$$\widehat{\mathbf{q}} = (\hat{q}_1, \ldots, \hat{q}_C)' \text{ where } \hat{q}_i = \sum_{\{s \in \mathcal{U}\}} I((A(\mathbf{s} \mid \mathcal{G}) = i))/N = v_i/N \tag{2.1}$$

where v_i is the number of observations in $\mathcal{U}$ classified by A to category i, and $N = \sum_i v_i$ is the sample size of $\mathcal{U}$. If $\mathcal{U}$ is large enough to be representative of the population in $\mathcal{D}_T$, it is clear that

$$\widehat{\mathbf{q}} \approx \mathbf{q} = (q_1, \ldots, q_C)' \text{ where } q_i = \int_{\mathcal{U}} P(A(\mathbf{s} \mid \mathcal{G}) = i \mid \mathbf{s}) dP(\mathbf{s}) = P(A(\mathbf{s}) = i),$$

i.e., $\widehat{\mathbf{q}}$ is the method-of-moments estimator of $\mathbf{q}$.

Unless the algorithm A trained on $\mathcal{D}_S$ perfectly agrees with the true membership assignment mechanism G in $\mathcal{D}_T$, there is no reason to consider $\mathbf{q}$ or $\widehat{\mathbf{q}}$ to be a good estimate of $\mathbf{p}$. More realistically, since $\mathcal{D}_S \neq \mathcal{D}_T$, accuracy depends on how similar the algorithm A is in the source and target domains. Hence, more generally we can think of $\mathbf{q}$ as the expected population class probabilities in $\mathcal{D}_T$ that would be predicted by $A(\cdot \mid \mathcal{G})$.

In their most general form, G and A can be thought of as measureable functions from the high-dimensional symptom space to the space of all C dimensional simplexes. Hence, we can write

$$A(\mathbf{s}) \sim \text{Multinomial}(\mathbf{q}),\ G(\mathbf{s}) \sim \text{Multinomial}(\mathbf{p})\,. \tag{2.2}$$

This only depicts the marginal distributions of $A(\mathbf{s})$ and $G(\mathbf{s})$. To infer about G from A, we need to model their joint distributions. We express $q_j = \sum_{i=1}^{C} m_{ij} p_i$ where $m_{ij} = p(A(\mathbf{s}) = j \mid G(\mathbf{s}) = i)$. In matrix notation, we have $\mathbf{q} = \mathbf{M}'\mathbf{p}$ where $\mathbf{M} = (m_{ij})$ is a transition matrix (i.e., $\mathbf{M1} = \mathbf{1}$) which we refer to as the *confusion matrix*. First note that, if $\mathbf{M} = \mathbf{I}$, then $\mathbf{p} = \mathbf{q}$ and hence this subsumes the case where class probabilities from the baseline algorithm is trusted as reliable surrogates of the true class probabilities.

For transfer learning to improve estimation of $\mathbf{p}$, we can opt to use the more general relationship $\mathbf{q} = \mathbf{M}'\mathbf{p}$ and estimate the transfer error rates m_{ij}'s from $\mathcal{L}$. Let $n = \sum_{i=1}^{C} n_i$ denote the sample size of $\mathcal{L}$ with n_i denoting the

number of objects belonging to class i. Also let

$$\mathbf{T} = (t_{ij}) = (\sum_{\mathbf{s}\in\mathcal{L}} I(A(\mathbf{s}) = j \mid G(\mathbf{s}) = i))$$

denote the *transfer error matrix* for algorithm A. Like many transfer learning algorithms, exploiting the transfer errors is key to our strategy. It is clear that t_{ij}/n_i is a method-of-moments estimator of m_{ij}.

We can use these estimates of m_{ij}, along with the earlier estimate of $\mathbf{q}$ to obtain a substantially improved estimate of $\mathbf{p}$. Formally we can specify this via a hierarchical model as:

$$\begin{array}{rl} A(\mathbf{s}_r) \overset{iid}{\sim} & \text{Multinomial}(1, \mathbf{M}'\mathbf{p}), r = 1, 2, \ldots, N \\ \mathbf{T}_{i*} \overset{ind}{\sim} & \text{Multinomial}(n_i, \mathbf{M}_{i*}), i = 1, 2, \ldots, C \end{array} \tag{2.3}$$

where for $r = 1, 2, \ldots, N$, $\mathbf{s}_r$ denote the covariate set for the r^{th} observation in $\mathcal{U}$, and for any matrix $\mathbf{M}$, $\mathbf{M}_{i*}$ and $\mathbf{M}_{*j}$ denote its i^{th} row and j^{th} column respectively. The top-row of (2.3) represents the relationship $\mathbf{q} = \mathbf{M}'\mathbf{p}$ and yields the method-of-moments estimators $\widehat{\mathbf{q}} = (v_1, v_2, \ldots, v_C)'/N$. The bottom-row of (2.3) is consistent with the naive estimates t_{ij}/n_i of m_{ij}.

To estimate $\mathbf{p}$, one can adopt a modular two-step approach where first $\widehat{\mathbf{q}}$ and $\widehat{\mathbf{M}}$ are calculated separately and then obtain

$$\widehat{\mathbf{p}} = \underset{\mathbf{p}:\mathbf{1}'\mathbf{p}=1,\ p_i\geq 0}{\arg\min}\ L(\widehat{\mathbf{q}}, \widehat{\mathbf{M}}'\mathbf{p})$$

where L is some loss function like the squared-error or, more appropriately, the Kullback-Liebler divergence between the probability vectors. This approach fails to propagate the uncertainty in the estimation of $\mathbf{M}$ in the final estimates of $\mathbf{p}$. Benefits of a one-stage approach over a two-stage one has

been demonstrated in recent work in transfer learning (Long et al., 2014). We recommend the one-stage information-theoretically optimal solution of estimating the joint MLE of $\mathbf{M}$ and $\mathbf{p}$ from (2.3).

The advantage of this simple transfer learning method is that it circumvents the need to improve the individual predictions of A in $\mathcal{D}_T$, and directly calibrates the population-level class probabilities $\mathbf{p}$, which are the quantities of interest here. We efficiently exploit the small local training data $\mathcal{L}$ to reduce cross-domain bias. Instead of trying to use $\mathcal{L}$ to estimate variants of a $S \times C$ matrix $\mathbf{P}(\mathbf{s} \mid c)$ describing propensities of manifestation of each symptom given each cause, as is used by many CCVA algorithms like Tariff, InSilicoVA etc., we now only use $\mathcal{L}$ to train a $(P(A(\mathbf{s}) \mid c))$ confusion matrix. Consequently, the matrix $(P(A(\mathbf{s}) \mid c))$ involves only $C(C-1)$ parameters as opposed to the $\mathcal{O}(SC)$ parameters of the $\mathbf{P}(\mathbf{s} \mid c)$ matrix. For verbal autopsy data, S is typically around 250 while we can choose C to be small focusing on the top $3-5$ causes. Hence, our approach achieves considerable dimension reduction by switching from the original covariate space to the predicted class space.

In equation (2.3) above, $\mathbf{q}$ can be estimated precisely because N is large. However, $\mathbf{M}$ has $C \times (C-1)$ parameters so that if there are many classes, the estimates of m_{ij} will have large variances owing to the small size of $\mathcal{L}$. Furthermore, in epidemiological studies, data collection often spans a few years; in the early stages, $\mathcal{L}$ may only have a very small sample size resulting in an extremely imprecise estimate of $\mathbf{M}$, even if we group the classes to a handful of larger classes. Consequently, in the next section we propose a

regularized approach that stabilizes the transfer learning.

2.2.2 Bayesian regularized approach

If $\mathcal{L}$ was not available, i.e., there is no labeled data in the target domain, we only have $\mathcal{U}$ and $\mathcal{G}$. Then it would be natural to train A using $\mathcal{G}$ and predict on $\mathcal{U}$ to obtain the estimates $\widehat{\mathbf{q}}$ as the best guess for $\mathbf{p}$. This is equivalent to setting $\mathbf{p} = \mathbf{q}$ and $\mathbf{M} = \mathbf{I}$, i.e., assuming that the algorithm A perfectly classifies in $\mathcal{D}_T$ even when trained only using $\mathcal{G}$ from $\mathcal{D}_S$. Extending this argument, when $\mathcal{L}$ is very small, direct estimates of $\mathbf{M}$ would be unstable and we should rely more on the predictions from A trained on $\mathcal{D}_S$. Hence, it is reasonable to shrink $\mathbf{p}$ towards $\mathbf{q}$ i.e., we shrink towards the default assumption that the baseline learner is accurate. This is equivalent to shrinking the estimate of $\mathbf{M}$ towards $\mathbf{I}$. The simplest way to achieve this is by using the regularized estimate $\widetilde{\mathbf{M}} = (1-\lambda)\widehat{\mathbf{M}} + \lambda\mathbf{I}$ where $\widehat{\mathbf{M}} = (\widehat{m}_{ij}) = t_{ij}/n_i$ is the unshrunk method-of-moments estimate of m_{ij} as derived in the previous section. The regularized estimate $\widetilde{\mathbf{M}}$ (like $\widehat{\mathbf{M}}$ and $\mathbf{M}$) remains a transition matrix. The parameter λ quantifies the degree of shrinkage with $\lambda = 0$ yielding the unbiased method-of-moments estimate and $\lambda = 1$ leading to $\widehat{\mathbf{p}} = \widehat{\mathbf{q}}$. Hence, λ represents the bias variance trade-off for estimation of transition matrices and for small sample sizes some intermediate values of λ may lead to better estimates of $\mathbf{M}$ and $\mathbf{p}$.

In epidemiological applications, as data will often come in batches over a period spanning few years, one needs to rerun the transfer learning procedure periodically to update the class probabilities. In the beginning, when $\mathcal{L}$ is extremely small, it is expected that more regularization is required. Eventually,

when $\mathcal{L}$ becomes large, we could rely on the direct estimate $\widehat{\mathbf{M}}$. Hence, λ should be a function of the size n of $\mathcal{L}$, with $\lambda = 1$ for $n = 0$ and $\lambda \approx 0$ for large n. Furthermore, at intermediate stages, since the distribution of true class memberships in $\mathcal{L}$ will be non-uniform across the classes, we will have a disparity in sample sizes n_i for estimating the different rows of $\mathbf{M}$. Consequently, it makes more sense to regularize each row of $\mathbf{M}$ separately instead of using a single λ. A more flexible regularized estimate is given by $\widetilde{\mathbf{M}}_{i*} = (1-\lambda_i)\widehat{\mathbf{M}}_{i*} + \lambda_i \mathbf{I}_{i*}$. The row specific weights λ_i should be chosen such that $\lambda_i = 1$ when $n_i = \sum_{j=1}^{C} t_{ij} = 0$, and $\lambda_i \approx 0$ when n_i is large. One choice to accomplish this is given by $\lambda_i = \gamma_i/(n_i + \gamma_i)$ for some fixed positive γ_i's.

We now propose a hierarchical Bayesian formulation that accomplishes this regularized estimation of any transition matrix $\mathbf{M}$. We consider a Dirichlet prior $\mathbf{M}_{i*} \overset{ind}{\sim} \text{Dirichlet}(\gamma_i(\mathbf{I}_{i*} + \epsilon\mathbf{1}))$ for the rows of $\mathbf{M}$. We first offer some heuristics expounding choice of this prior. We will have $\mathbf{M}_{i*} \mid \mathbf{T}_{i*}, \gamma_i \sim \text{Dirichlet}(\mathbf{T}_{i*} + \gamma_i(\mathbf{I}_{i*} + \epsilon 1))$. Hence,

$$E(\mathbf{M}_{i*} \mid \mathbf{T}_{i*}, \gamma_i) = \frac{\mathbf{T}_{i*} + \gamma_i(\mathbf{I}_{i*} + \epsilon\mathbf{1})}{n_i + \gamma_i(1 + C\epsilon)} \overset{\epsilon \to 0}{\to} (1-\lambda_i)\frac{\mathbf{T}_{i*}}{n_i} + \lambda_i \mathbf{I}_{i*} \text{ where } \lambda_i = \frac{\gamma_i}{n_i + \gamma_i}.$$

Hence, using a small enough ϵ, the Bayes estimator (posterior mean) for $\mathbf{M}$ becomes equivalent with the desired shrinkage estimator $\widetilde{\mathbf{M}}_{i*}$ proposed above. When $n = 0$, the Bayes estimate $E(\mathbf{M} \mid \mathbf{T}, \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_C)') \approx \mathbf{I}$, and for large n, $E(\mathbf{M} \mid \mathbf{T}, \boldsymbol{\gamma})$ becomes the method-of-moments estimator $\widehat{\mathbf{M}}$. Hence, the Dirichlet prior ensures that in data-scarce setting, $\mathbf{M}$ is shrunk towards $\mathbf{I}$ and consequently $\mathbf{p}$ towards $\mathbf{q}$. We note that however this initial exposition for the posterior of $\mathbf{p}$ are derived conditional on estimation of $\mathbf{M}$

as an independent piece and ignores the data from $\mathcal{U}$. In Theorem 1, we will present a more formal result that looks at the properties of the marginal posterior of $\mathbf{p}$.

To complete the hierarchical formulation, we augment (2.3) with the priors:

$$\begin{aligned} \mathbf{M}_{i*} &\overset{ind}{\sim} \text{Dirichlet}(\gamma_i(\mathbf{I}_{i*} + \epsilon\mathbf{1})), i = 1, 2, \ldots, C \\ \mathbf{p} &\sim \text{Dirichlet}(\delta\mathbf{1}) \\ \gamma_i &\overset{ind}{\sim} Gamma(\alpha, \beta), i = 1, 2, \ldots, C \end{aligned} \tag{2.4}$$

In practice, we need to use a small $\epsilon > 0$ to ensure a proper posterior for $\mathbf{M}$ when any off-diagonal entries of $\mathbf{T}$ are zero, which is very likely due to the limited size of $\mathcal{L}$. Note that our model only uses the data from $\mathcal{L}$ to estimate the conditional probabilities $P(A(\mathbf{s}) \mid G(\mathbf{s}))$ for $\mathbf{s} \in \mathcal{L}$. We do not model the marginal distribution of $A(\mathbf{s})$ for $\mathbf{s} \in \mathcal{L}$ like we do for $\mathbf{s}$ in $\mathcal{U}$. This is because often data for the labeled set is collected under controlled settings, and marginal distribution of the covariates for the samples in $\mathcal{L}$ is not representative of the true marginal distribution of the covariates in $\mathcal{D}_T$. Hence, we only use $\mathcal{L}$ to estimate the conditional probabilities $\mathbf{M}$.

Our previous heuristic arguments, illustrating the shrinkage estimation of $\mathbf{M}$ induced by the Dirichlet prior, are limited to the estimation of $\mathbf{M}$ from $\mathcal{L}$ as an independent piece and disregards the data and model for $\mathcal{U}$, i.e. the first row of (2.3). In a hierarchical setup, however, the models for $\mathcal{U}$ and $\mathcal{L}$ contribute jointly to the estimation of $\mathbf{M}$ and $\mathbf{p}$. We will now state a more general result that argues that for our full hierarchical model specified through (2.3) and (2.4), when there is no labeled data in $\mathcal{D}_T$ or if the algorithm A demonstrates perfect accuracy (zero transfer error) on $\mathcal{L}$, then

the marginal posterior estimates of $\mathbf{p}$ from our model coincides with the baseline estimates $\widehat{\mathbf{q}}$. Before stating the result, first note that the likelihood for $\mathbf{a} = (A(\mathbf{s}_1), A(\mathbf{s}_2), \ldots, A(\mathbf{s}_N))'$ can be represented using the sufficient statistics $\mathbf{v} = (v_1, v_2, \ldots, v_C)'$. We can write $p(\mathbf{a}) \propto \prod_{j=1}^{C} q_j^{v_j}$ and hence $\mathbf{p}, \mathbf{M}, \boldsymbol{\gamma} | \text{data} = \mathbf{p}, \mathbf{M}, \boldsymbol{\gamma} \mid \mathbf{v}, \mathbf{T}$.

Theorem 1. *If* $\mathbf{T}$ *is a diagonal matrix, i.e., either there is no* $\mathcal{L}$*, or* A *classifies perfectly on* $\mathcal{L}$*, then* $\lim_{\epsilon \to 0} \mathbf{p} \mid \mathbf{v}, \mathbf{T} \sim \textit{Dirichlet}(\mathbf{v} + \delta \mathbf{1})$*. For* $\delta = 0$*,* $\lim_{\epsilon \to 0} E(\mathbf{p} \mid \mathbf{v}, \mathbf{T}) = \widehat{\mathbf{q}}$.

Note that Theorem 1 is a result about the posterior of our quantity of interest $\mathbf{p}$, marginalizing out the other parameters $\mathbf{M}$, and the γ_i's from the hierarchical model specified through equations (2.3) and (2.4). We also highlight that this is not an asymptotic result and holds true for any sample size, as long as we choose ϵ and δ to be small. This is important as our manuscript pertains to epidemiological applications where $\mathcal{L}$ will be extremely small and asymptotic results are not relevant.

Theorem 1 also does not require any assumption about the underlying data generation scheme, and is simply a desirable property of our transfer learning model. If there is no labeled data in $\mathcal{D}_T$, then it is natural to trust the $P(c \mid \mathbf{s})$ map learnt by A on a source domain and only learn the target domain marginal distributions of $\mathbf{s}$ from $\mathcal{U}$ to arrive at the estimates $\widehat{\mathbf{q}}$ of $\mathbf{p}$. Similarly, in the best case scenario, when A is absolutely accurate for the target domain, Theorem 1 guarantees that our model automatically recognizes this accuracy and does not modify the baseline estimates $\widehat{\mathbf{q}}$ from A. The result of Theorem 1 is confirmed in simulations in Section 2.5.

Although Theorem 1 is assumption-free, it only concerns with the performance of the model when there is no $\mathcal{L}$ or when A is perfect on $\mathcal{L}$. While this is a good sanity check for our model, realistically we will have a small $\mathcal{L}$ where A will be inaccurate. In such cases, the performance of our model will of course depend on the data generation process. Hence, we summarize the data generation assumption that drive the model formulation. Since, there is no labeled data in $\mathcal{U}$, we need to assume some commonality between $\mathcal{L}$ and $\mathcal{U}$ in order for the labeled data in $\mathcal{L}$ to be useful for estimating the CSMFs in $\mathcal{U}$. Hence, the model assumes that the conditional distribution of $A(\mathbf{s}) \mid G(\mathbf{s})$ (i.e., the $\mathbf{M}$ matrix) is same in $\mathcal{U}$ and $\mathcal{L}$. We would like to emphasize that we do not assume that the marginal distributions of the symptoms $\mathbf{s}$ or the cause $G(\mathbf{s})$ (i.e., the CSMFs) are same in any of $\mathcal{G}$, $\mathcal{U}$ and $\mathcal{L}$. Of course, the assumption of same confusion matrix $\mathbf{M}$ for $\mathcal{U}$ and $\mathcal{L}$ can also be incorrect (all models are wrong). However, the class of models spanned by use of a general $\mathbf{M}$ is a superset of the default approach of using the baseline classifier (i.e., assuming $\mathbf{M} = \mathbf{I}$). Also, we can relax the assumption of constant $\mathbf{M}$ between $\mathcal{U}$ and $\mathcal{L}$ to make entries of $\mathbf{M}$ function of some covariates. This model and its implementation is discussed in Section 2.4. This would lead to substantial increase in parameter dimensionality and is only recommended when $\mathcal{L}$ is large.

2.2.3 Gibbs sampler using augmented data

We devise an efficient implementation of the hierarchical transfer learning model using a data augmented Gibbs sampler. The joint posterior density can

be expressed as

$$p(\mathbf{p}, \mathbf{M}, \gamma \mid \mathbf{v}, \mathbf{T}) \propto \; p(\mathbf{v} \mid \mathbf{M}, \mathbf{p}) p(\mathbf{T} \mid \mathbf{M}) p(\mathbf{M} \mid \gamma) p(\mathbf{p}) p(\gamma)$$

Let $\mathbf{p} \mid \cdot$ denote the full conditional distribution of $\mathbf{p}$. We use similar notation for other full conditionals. First note that since $p(\mathbf{v} \mid \mathbf{M}, \mathbf{p}) \propto \prod_j (\sum_i m_{ij} p_i)^{v_j}$, the full conditional densities $\mathbf{p} \mid \cdot$ and $\mathbf{M} \mid \cdot$ do not belong to any standard family of distributions, thereby prohibiting a direct Gibbs sampler. We here use a data augmentation scheme enabling a Gibbs sampler using conjugate distributions.

The term $(\sum_i m_{ij} p_i)^{v_j}$ can be expanded using the multinomial theorem, with each term corresponding to one of the partitions of v_j into C non-negative integers. Equivalently we can write

$$(\sum_i m_{ij} p_i)^{v_j} \propto E(\prod_i (m_{ij} p_i)^{b_{ij}}) \text{ where } \mathbf{b}_j = (b_{1j}, \ldots, b_{Cj})' \sim \text{ Multinomial}(v_j, \mathbf{1}/C).$$

Choosing $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_C$ to be independent, we can express $\prod_j (\sum_i m_{ij} p_i)^{v_j} \propto E(\prod_j \prod_i (m_{ij} p_i)^{b_{ij}})$ where the proportionality constant only depends on the observed v_j's. Using the augmented data matrix $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_C) = (b_{ij})$, we can write the complete posterior as

$$\begin{aligned} p(\mathbf{p}, \mathbf{M}, \gamma, \mathbf{B} \mid \mathbf{v}, \mathbf{T}) \propto \; & \prod_j \prod_i (m_{ij} p_i)^{b_{ij}} \times \prod_i p_i^{\delta - 1} \times \prod_i \gamma_i^{\alpha - 1} \exp(-\beta \gamma_i) \times \\ & \prod_i \left(\frac{\Gamma(C\gamma_i \epsilon + \gamma_i)}{\Gamma(\gamma_i \epsilon)^{C-1} \Gamma(\gamma_i \epsilon + \gamma_i)} \prod_j (m_{ij})^{t_{ij} + \gamma_i \epsilon + \gamma_i 1(i=j) - 1} \right) \end{aligned} \tag{2.5}$$

The full conditional distributions can be updated as follows (derivations

omitted):

$$\mathbf{b}_j \mid \cdot \sim Multinomial(v_j, \frac{1}{\sum_i m_{ij}p_i}(m_{1j}p_1, m_{2j}p_2, \ldots, m_{Cj}p_C)')$$

$$\mathbf{M}_{i*} \mid \cdot \sim Dirichlet(b_{i1} + \gamma_i\epsilon + t_{i1}, \ldots, b_{ii} + \gamma_i\epsilon + t_{ii} + \gamma_i, \ldots, b_{iC} + \gamma_i\epsilon + t_{iC})$$

$$\mathbf{p} \mid \cdot \sim Dirichlet(\sum_j b_{1j} + \delta, \ldots, \sum_j b_{Cj} + \delta)$$

$$p(\gamma_i \mid \cdot) \propto \frac{\Gamma(C\gamma_i\epsilon + \gamma_i)}{\Gamma(\gamma_i\epsilon)^{C-1}\Gamma(\gamma_i\epsilon + \gamma_i)}\gamma_i^{\alpha-1}\exp(-\beta\gamma_i)\prod_j m_{ij}^{\gamma_i\epsilon+\gamma_i 1(i=j)}$$

The data augmentation ensures that, except the C γ_i's, which are updated using a metropolis random walk with log-normal proposals, all the other $\mathcal{O}(C^2)$ parameters are update by sampling from standard distributions leading to an extremely fast and efficient Gibbs sampler.

2.3 Ensemble transfer learning

Let there be K classifiers $A^{(1)}, A^{(2)}, \ldots, A^{(K)}$ and let $\mathbf{a}^{(k)} = (a_1^{(k)}, a_2^{(k)}, \ldots, a_N^{(k)})'$ be the predicted class memberships from the k^{th} algorithm for all the N observations in $\mathcal{U}$. Let $\mathbf{v}^{(k)}$ denote the vector of counts of predicted class memberships on $\mathcal{U}$ using $A^{(k)}$. We expect variation among the predictions from the different classifiers and consequently among the baseline estimates of population-level class probabilities $\widehat{\mathbf{q}}^{(k)} = \mathbf{v}^{(k)}/N$ and their population equivalents $\mathbf{q}^{(k)} = P(A^{(k)}(\mathbf{s}))$. Since the true population class probability vector $\mathbf{p}$ is unique, following Section 2.2.1 we can write $\mathbf{q}^{(k)} = (q_1^{(k)}, q_2^{(k)}, \ldots, q_C^{(k)})' = \mathbf{M}^{(k)'}\mathbf{p}$ where $\mathbf{M}^{(k)} = (m_{ij}^{(k)})$ is now the classifier-specific confusion matrix. The predicted class membership for the r^{th} observation in $\mathcal{U}$ by algorithm $A^{(k)}$,

denoted by $a_r^{(k)}$, marginally follows a Multinomial$(1, \mathbf{q}^{(k)})$ distribution. We have K such predictions for the same observation, one for each classifier, and these are expected to be correlated. So, we need to look at the joint distribution of the K C-dimensional multinomial random variables. Since, in its most general form this will involve $\mathcal{O}(C^K)$ parameters, we use a pragmatic simplifying assumption to derive the joint distribution. We assume that $a_r^{(1)}, a_r^{(2)}, \ldots, a_r^{(K)}$ are independent conditional on $G(\mathbf{s}_r)$, i.e.

$$p(a_r^{(1)} = j_1, a_r^{(2)} = j_2, \ldots, a_r^{(K)} = j_K \mid G(\mathbf{s}_r) = i) = \prod_{k=1}^{K} m_{ij_k}^{(k)} \tag{2.6}$$

This assumption is unlikely to hold in reality but is a common dimension reducing assumption used in classification problems. For example, the naive Bayes classifier uses this assumption to jointly model the probability of covariates given the true class memberships. Similar assumptions are used by InSilicoVA and InterVA to derive the joint distribution of the vector of symptoms $\mathbf{s}_r$. Here we are applying the same assumption but not on $\mathbf{s}_r$ but on the lower-dimensional prediction vector $(a_r^{(1)}, a_r^{(2)}, \ldots, a_r^{(K)})'$.

Under this assumption, the marginal independence of the $a_r^{(k)}$'s will not generally hold. Instead we will have

$$p(\mathbf{a}_r = \mathbf{j}) = p(a_r^{(1)} = j_1, a_r^{(2)} = j_2, \ldots, a_r^{(K)} = j_K) = \sum_{i=1}^{C} \left(\prod_{k=1}^{K} m_{ij_k}^{(k)} \right) p_i = w_{\mathbf{j}} \tag{2.7}$$

where $\mathbf{j} = (j_1, j_2, \ldots, j_K)$ denotes a $C \times 1$ vector index.

From the limited labeled dataset $\mathcal{L}$ in the target domain $\mathcal{D}_T$, the classifier specific transfer error matrices $\mathbf{T}^{(k)} = (t_{ij}^{(k)}) = (\sum_{\mathbf{s} \in \mathcal{L}} I(A^{(k)}(\mathbf{s}) = j \mid G(\mathbf{s}) = i))$ are also known and can be used to estimate the respective confusion

matrices $\mathbf{M}^{(k)}$ in the same way $\mathbf{M}$ was estimated from $\mathbf{T}$ in Section 2.2.1. To introduce shrinkage in the estimation of $\mathbf{M}^{(k)}$, like in Section 2.2.2, we assign Dirichlet priors for each $\mathbf{M}^{(k)}$.

Let $\mathbf{w}$ denote a $C^K \times 1$ vector formed by stacking up all the $q_{j_1,j_2,\ldots,j_K}$'s defined in (2.7). The full specifications for the ensemble model that incorporates the predictions from all the algorithms is given by:

$$\begin{aligned}
\mathbf{a}_r &\overset{iid}{\sim} \text{Multinomial}(1, \mathbf{w}), r = 1, 2, \ldots, N \\
\mathbf{T}^{(k)}_{i*} &\overset{ind}{\sim} \text{Multinomial}(n_i, \mathbf{M}^{(k)}_{i*}), i = 1, 2, \ldots, C;\ k = 1, 2, \ldots, K \\
\mathbf{M}^{(k)}_{i*} &\overset{ind}{\sim} \text{Dirichlet}(\gamma^{(k)}_i(\mathbf{I}_{i*} + \epsilon\mathbf{1})), i = 1, 2, \ldots, C;\ k = 1, 2, \ldots, K \\
\mathbf{p} &\sim \text{Dirichlet}(\delta\mathbf{1}) \\
\gamma^{(k)}_i &\overset{ind}{\sim} Gamma(\alpha, \beta), i = 1, 2, \ldots, C;\ k = 1, 2, \ldots, K
\end{aligned} \tag{2.8}$$

Although $\mathbf{w}$ is a $C^K \times 1$ vector, courtesy of the conditional independence assumption (2.6), it is only parameterized using the matrices $\mathbf{M}^{(k)}$ and $\mathbf{p}$ as specified in (2.7), and hence involves $KC^2 + C$ parameters. This ensures that there is adequate data to estimate the enhanced number of parameters for this ensemble method, as for each $\mathbf{M}^{(k)}$ we observe the corrsponding trasfer error matrix $\mathbf{T}^{(k)}$. The Gibbs sampler for (2.8) is provided in Section 2.9.3. To understand how the different classifiers are given importance based on their transfer errors on $\mathcal{L}$, we present the following result:

Theorem 2. *If $\mathbf{T}^{(1)}$ is diagonal with positive diagonal entries, and all entries of $\mathbf{T}^{(k)}$ are ≥ 1 for all $k > 1$, then $\mathbf{p} \mid data \sim Dirichlet\,(\mathbf{v}^{(1)} + \delta)$. For $\delta = 0$, $E(\mathbf{p} \mid data) = \mathbf{q}^{(1)}$.*

Theorem 2 reveals that if one of the K algorithms (which we assume to be the first algorithm without loss of generality) produce perfect prediction

on $\mathcal{L}$, then posterior mean estimate of $\mathbf{p}$ from the ensemble model coincides with that of the baseline estimate from that classifier. The perfect agreement assumed in Theorem 2 will not occur in practice. However, simulation and data analyses will confirm that the estimate of $\mathbf{p}$ from the ensemble model tend to agree with that from the single-classifier model in Section 2.2.2 with the more accurate algorithm. This offers a more efficient way to weight the multiple algorithms, yielding a unified estimate of class probabilities that is more robust to inclusion of an inaccurate algorithm in the decision making. In comparison, a simple average of estimated $\mathbf{p}$'s from single-classifier transfer learning models for each of the K algorithms would be more adversely affected by inaccurate algorithms.

2.3.1 Independent ensemble model

The likelihood for the top-row of (2.8) is proportional to $\prod_{\mathbf{j}} w_{\mathbf{j}}^{y_{\mathbf{j}}}$ where $y_{\mathbf{j}} = \sum_{\mathbf{s}\in\mathcal{U}} I(a^{(1)} = j_1, \ldots, a^{(K)}(\mathbf{s}) = j_K)$ denote the total number of observations in $\mathcal{U}$ where the predicted class-memberships from the K algorithms corresponds to the combination $\mathbf{j} = (j_1, \ldots, j_K)'$. Even though $\mathcal{U}$ will be moderately large (few thousand observations in most epidemiological applications), unless both C and K are very small ($C \leq 5$ and $K \leq 3$), $y_{\mathbf{j}}$'s will be zero for most of the C^K possible combinations $\mathbf{j}$. This will in-turn affect the estimates of $\mathbf{w}$. For applications to verbal autopsy based estimation of population CSMFs, there are many CCVA algorithms (as introduced in Section 2.1), and researchers often want to use all of them in an analysis. We also may be interested in more than $3-5$ top causes. In such cases, the extremely sparse C^K vector

formed by stacking up the $\mathbf{y_j}$'s will destabilize the estimation of $\mathbf{w}$. Also, the Gibbs sampler (see Section 2.9.3) of the joint-ensemble model introduces an additional C^K independent multinomial variables of dimension C thereby accruing substantial computational overhead and entailing long runs of the high-dimensional Markov chain to achieve convergence.

In this section, we offer a pragmatic alternative model for ensemble transfer learning that is computationally less demanding. From equation (2.7), we note that

$$p(a_r^{(k)} = j_k) = \sum_{j_s : s \neq k} \sum_{i=1}^{C} \left(\prod_{k=1}^{K} m_{ij_k}^{(k)} \right) p_i = \sum_{i=1}^{C} m_{ij_k}^{(k)} p_i \tag{2.9}$$

by exchanging the summations. Hence, the marginal distribution of $a_r^{(k)}$ is Multinomial$(1, \mathbf{q}^{(k)})$ where $\mathbf{q}^{(k)} = (\mathbf{M}^{(k)})'\mathbf{p}$. We model the $a_r^{(k)}$'s independently for each k, ignoring the correlation among the predictions in $\mathcal{U}$ from the K classifiers as follows:

$$\mathbf{a}_r = (a_r^{(1)}, a_r^{(2)}, \ldots, a_r^{(K)})' \overset{iid}{\sim} \prod_{k=1}^{K} \text{Multinomial}(1, \mathbf{q}^{(k)}), r = 1, \ldots, N \tag{2.10}$$

We replace the top-row of (2.8) with (2.10), keeping the other specification same as in (2.8). We call this the independent ensemble model. Note that, while we only use the marginal distributions of the $a_r^{(k)}$'s ignoring their joint dependence, the joint distribution is preserved in the model for the transfer errors on $\mathcal{L}$ specified in the second-row of (2.8), as all the $\mathbf{M}^{(k)}$'s are tied to the common truth $\mathbf{p}$ through the equations $\mathbf{q}^{(k)} = \mathbf{M}^{(k)'}\mathbf{p}$. While the total number of parameters for the joint and independent ensemble models remain the same, eliminating the joint model for each of the C^K combination of predicted causes from the K algorithms allows decomposing the likelihood for (2.10) as product

of individual likelihoods on $\mathcal{U}$ for each of the K classifiers. Additionally, the Gibbs sampler for the independent ensemble model is much simpler and closely resembles the sampler for the single-classifier model in Section 2.2.3. We only need to introduce K $C \times C$ matrices $\mathbf{B}^{(k)} = (\mathbf{b}_1^{(k)}, \mathbf{b}_2^{(k)}, \ldots, \mathbf{b}_C^{(k)})$, one corresponding to each CCVA algorithm, akin to the matrix $\mathbf{B}$ introduced in Section 2.2.3. The Gibbs sampler steps for the independent ensemble model are:

$$\mathbf{b}_j^{(k)} \mid \cdot \sim Multinomial(v_j^{(k)}, \frac{1}{\sum_i m_{ij}^{(k)} p_i}(m_{1j}^{(k)} p_1, m_{2j}^{(k)} p_2, \ldots m_{Cj}^{(k)} p_C)')$$

$$\mathbf{M}_{i*}^{(k)} \mid \cdot \sim Dirichlet(\mathbf{B}_{i*}^{(k)} + \mathbf{T}_{i*}^{(k)} + \gamma_i^{(k)} \mathbf{I}_{i*} + \epsilon \boldsymbol{\gamma}_i^{(k)} \mathbf{1})$$

$$\mathbf{p} \mid \cdot \sim Dirichlet(\sum_k \sum_j b_{1j}^{(k)} + \delta, \ldots, \sum_k \sum_j b_{Cj}^{(k)} + \delta)$$

Observe that the sampler for the independent model uses CK additional parameters as opposed to C^K parameters introduced in the joint sampler. This ensures that the MCMC dimensionality does not exponentially increase if predictions from more algorithms are included in the ensemble model. The theoretical result in Theorem 2 no longer remains true for the independent model. However, our simulation results in Section 2.9.5.5 of the supplementary material (http://www.biostatistics.oxfordjournals.org) show that in practice it continues to put higher weights on the more accurate algorithm and consistently performs similar to or better than the joint model.

2.4 Demographic covariates and spatial information

The transfer-learning model introduced up to this point is focused on generating population-level estimates of the CSMF $\mathbf{p}$. An important extension for epidemiological applications would be to model $\mathbf{p}$ as a function of covariates like geographic region, social economic status (SES), sex and age groups. This will enable the estimation of regional and age-sex stratified estimates. In this section, we generalize the model to accommodate covariates. We illustrate for the single-classifier model in Section 2.2.2; a similar approach extends the ensemble model.

Let $\mathbf{x}_r$ denote a vector of covariates for the r^{th} VA record in $\mathcal{U}$. We propose the following modifications to the model for allowing covariate-specific class distributions $\mathbf{p}_r = (p_{r1}, p_{r2}, \ldots, p_{rC})'$:

$$\begin{aligned} A(\mathbf{s}_r) &\overset{ind}{\sim} \text{Multinomial}(\mathbf{M}'\mathbf{p}_r), r = 1, 2, \ldots, N \\ p_{ri} &= \frac{exp(\mathbf{x}_r'\boldsymbol{\beta}_i)}{\sum_{i=1}^{C} exp(\mathbf{x}_r'\boldsymbol{\beta}_i)}, i = 1, 2, \ldots, C, \boldsymbol{\beta}_C = 0 \\ \boldsymbol{\beta}_i &\overset{ind}{\sim} N(\mathbf{m}_{0i}, \mathbf{W}_{0i}) \end{aligned} \tag{2.11}$$

All other components of the original model in (2.3) and (2.4) remain unchanged. The middle row of (2.11) specify a multi-logistic model for the class probabilities using the covariates. The top row uses the covariate specific $\mathbf{p}_r$ to model the analogous class probabilities $\mathbf{q}_r = \mathbf{M}'\mathbf{p}_r$ as would be predicted by A. Finally, the bottom row specifies Normal priors for the regression coefficients. The switch from a Dirichlet prior for $\mathbf{p}$ to the multi-logistic model implies we can no longer directly leverage conjugacy in the Gibbs sampler. Polson, Scott, and Windle, 2013 proposed a Polya-Gamma data augmentation scheme to allow conjugate sampling for generalized linear models. We now

show how our own data augmentation scheme introduced in Section 2.2.3 harmonizes with the Polya-Gamma sampler to create a streamlined Gibbs sampler.

2.4.1 Gibbs sampler using Polya-Gamma scheme

We will assume there are G unique combinations of covariate values – for example, if there are four geographic regions and three age groups, then $G = 12$. If we have a continuous covariate, then $G = N$, where N is the number of subjects sampled in $\mathcal{U}$. Then letting g, $g = 1, \ldots, G$, represent a specific covariate combination $\mathbf{x}_g$, we can again represent the likelihood for $\mathbf{a} = (A(\mathbf{s}_1), A(\mathbf{s}_2), \ldots, A(\mathbf{s}_N))'$ using the $G \times C$ sufficient statistics $\mathbf{V} = (v_{gj})$ where v_{gi} is the total number of subjects with covariate values g that were predicted to have died of cause i. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{C-1})$. We now have

$$p(\mathbf{V} \mid \mathbf{M}, \boldsymbol{\beta}) \propto \prod_{g=1}^{G} \prod_{j=1}^{C} (\sum_{i=1}^{C} m_{ij} p_{gi})^{v_{gj}}$$

and the joint posterior density can now be expressed as

$$p(\boldsymbol{\beta}, \mathbf{M}, \boldsymbol{\gamma} \mid \mathbf{V}, \mathbf{T}) \propto \; p(\mathbf{V} \mid \mathbf{M}, \boldsymbol{\beta}) p(\mathbf{T} \mid \mathbf{M}) p(\mathbf{M} \mid \boldsymbol{\gamma}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma})$$

The terms that are different from Section 2.2.3 are $p(\mathbf{V} \mid \mathbf{M}, \boldsymbol{\beta})$ and $p(\boldsymbol{\beta})$. The sampling step for $\boldsymbol{\gamma}$ remains exactly the same as previously discussed. We will use a similar data augmentation strategy as in Section 2.2.3 and combine with a Polya-Gamma data augmentation to sample from this posterior distribution. We expand the term $(\sum_i m_{ij} p_{gi})^{v_{gj}} \propto E(\prod_i (m_{ij} p_{gi})^{b_{gij}})$ where

$$\mathbf{b}_{gj} = (b_{g1j}, \ldots, b_{gCj})' \overset{ind}{\sim} \text{Multinomial}(v_{gj}, \mathbf{1}/C).$$

Let $\mathbf{B}$ denote the $GC \times C$ matrix formed by stacking all the $\mathbf{b}_{gj}$'s row-wise. We can write

$$p(\boldsymbol{\beta}, \mathbf{B}, \mathbf{M}, \boldsymbol{\gamma} \mid \mathbf{V}, \mathbf{T}) \propto \prod_g \prod_i \prod_j (m_{ij} p_{gi})^{b_{gij}} \times p(\mathbf{T} \mid \mathbf{M}) p(\mathbf{M} \mid \boldsymbol{\gamma}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma})$$

The following updates ensue immediately:

$$\mathbf{b}_{gj} \mid \cdot \sim Multinomial(v_{gj}, \frac{1}{\sum_i m_{ij} p_{gi}} (m_{1j} p_{g1}, m_{2j} p_{g2}, \ldots, m_{Cj} p_{gC})')$$

$$\mathbf{M}_{i*} \mid \cdot \sim Dirichlet \left(\mathbf{T}_{i*} + \boldsymbol{\gamma}_i \mathbf{I}_{i*} + \boldsymbol{\gamma}_i \mathbf{1} + (\sum_g b_{gi1}, \ldots, \sum_g b_{giC})' \right)$$

For $\boldsymbol{\beta}_i$'s we introduce the Polya-Gamma variables ω_{gi}'s and define $\boldsymbol{\Omega}_i = \text{diag}(\{\omega_{gi}\}_{g=1}^G)$, $n_g = \sum_j v_{gj}$, and $\boldsymbol{\kappa}_i = (\kappa_{1i}, \ldots, \kappa_{Gi})'$ where $\kappa_{gi} = \sum_j b_{gij} - n_g/2$. Defining $\mathbf{W}_i^{-1} = \mathbf{X}' \boldsymbol{\Omega}_i \mathbf{X} + \mathbf{W}_{0i}^{-1}$, we then have

$$\omega_{gi} \mid \cdot \sim PG(n_g, \mathbf{x}_g^T \boldsymbol{\beta}_i - c_{gi}) \text{ where } c_{gi} = \log(\sum_{k \neq i} exp(x_g^T \beta_k))$$

$$\boldsymbol{\beta}_i \mid \cdot \sim \mathcal{N}(\mathbf{m}_i, \mathbf{W}_i) \text{ where } \mathbf{m}_i = \mathbf{W}_i \left(\mathbf{X}'(\boldsymbol{\kappa}_i - \boldsymbol{\Omega}_i \mathbf{c}_i) + \mathbf{W}_{0i}^{-1} \mathbf{m}_{0i} \right)$$

Here PG denotes the Polya-Gamma distribution and $\mathbf{c}_i = (c_{1i}, c_{2i}, \ldots, c_{Gi})'$. This completes the steps of a Gibbs sampler where all the parameters except $\boldsymbol{\gamma}$ are updated via sampling from conjugate distributions. We can transform the posterior samples of $\boldsymbol{\beta}$ to obtain posterior samples of p_{gi}. Estimates of the marginal class distribution for the whole country can also be obtained by using the relationship $p_i = \int p_{gi} dP(g)$ where an empirical estimate of the covariate distribution $P(g)$ can be obtained from $\mathcal{U}$.

2.4.2 Covariate-specific transfer error

Until now, we have assumed that the transition matrix $\mathbf{M}$ is independent of the covariates. We can also introduce covariates in modeling the conditional probabilities m_{ij}'s using a similar multi-logistic regression. This model will be particularly useful if there is prior knowledge about covariate-dependent biases in the predictions from a classifier. Letting m_{rij} denoting the conditional probabilities $p(A(\mathbf{s}) = j \mid G(\mathbf{s}) = i, \mathbf{x}_r)$ we can model

$$\begin{aligned} m_{rij} = &\ \frac{exp(\mathbf{x}_r'\boldsymbol{\zeta}_{ij})}{\sum_{i=1}^{C} exp(\mathbf{x}_r'\boldsymbol{\zeta}_{ij})}, i, j \in \{1, 2, \ldots, C\}, \boldsymbol{\zeta}_{iC} = 0 \\ \boldsymbol{\zeta}_{ij} \overset{ind}{\sim} &\ N(\mathbf{m}_{0ij}, \mathbf{W}_{0ij}),\ j < C\ . \end{aligned} \tag{2.12}$$

The implementation will involve Polya-Gamma samplers for each row of $\mathbf{M}$ in a manner exactly similar to the sampler outlined above (we omit the details). Since we can only estimate the parameters $\boldsymbol{\zeta}_{ij}$ from the limited local data, we can only adopt this approach with a very small set of covariates for modeling the transfer error rates.

2.5 Simulation studies

The Population Health Metrics Research Consortium (PHMRC) study, conducted in 4 countries across six sites, is a benchmark database of paired VA records and GS-COD of children, neonates and adults. PHMRC data is frequently used to assess performance of CCVA algorithms. We conduct a set of simulation studies using the PHMRC data (obtained through the openVA package, version 1.0.5) to generate a wide range of plausible scenarios where the performance of of our transfer learning models needs to be assessed with

Table 2.2: List of models used to estimate population CSMF

Model name	Description
Tariff$_{\mathcal{G}}$	Tariff trained on the source-domain gold standard data $\mathcal{G}$
Tariff$_{BTL}$	Bayesian transfer learner using the output from Tariff$_{\mathcal{G}}$
InSilico$_{\mathcal{G}}$	InSilicoVA trained on the source-domain gold standard data $\mathcal{G}$
InSilicoVA$_{BTL}$	Bayesian transfer learner using the output from InSilicoVA$_{\mathcal{G}}$
Ensemble$_{I}$	Ensemble Bayesian transfer learner (independent) using Tariff$_{\mathcal{G}}$ and Insilico$_{\mathcal{G}}$

respect to the popular CCVA algorithms. First, we randomly split the PHMRC child data (2064 samples) into three parts representing $\mathcal{G}$, and initial $\mathcal{L}$ and initial $\mathcal{U}$ respectively using a 2:1:2 ratio, containing roughly 800, 400 and 800 samples respectively. As accurate estimation of mortality fractions from most prevalent causes are usually the priority, we restrict our attention to four causes: the top three most prevalent causes in the target domain data ($\mathcal{L} \cup \mathcal{U}$) – Pneumonia, Diarrhea/Dysentry, Sepsis, and an *Other* cause grouping together all the remaining causes.

We wanted to simulate scenarios where both a) the marginal distributions $P(c) = P(G(\mathbf{s}) = c)$ of the classes, and b) the conditional distributions $P(c \mid \mathbf{s})$ are different between the source and target domains. To ensure the latter, given a confusion matrix $\mathbf{M} = (m_{ij})$ we want $P(A(\mathbf{s}) = j \mid G(\mathbf{s}) = i) = m_{ij}$ for any $\mathbf{s} \in \mathcal{L} \cup \mathcal{U}$. We will achieve this by discarding the actual labels in $\mathcal{L} \cup \mathcal{U}$ and generating new labels such that an algorithm A trained on $\mathcal{G}$ shows transfer error rates quantified by $\mathbf{M}$ on $\mathcal{L} \cup \mathcal{U}$. Additionally, the new labels need to be assigned in a way to ensure that the target domain class probability vector is $\mathbf{p}_{\mathcal{U}}$, for any choice of $\mathbf{p}_{\mathcal{U}}$ different from the source domain class probabilities in $\mathbf{p}_{\mathcal{G}}$.

Note that if the true population class probabilities in $\mathcal{D}_T$ needs to be $\mathbf{p}_{\mathcal{U}}$,

then $\mathbf{q}_{\mathcal{U}}$, the population class probabilities as predicted by A is given by $\mathbf{q}_{\mathcal{U}} = \mathbf{M}'\mathbf{p}_{\mathcal{U}}$. Hence, we first use A trained on $\mathcal{G}$ to predict the labels for each $\mathbf{s}$ in the initial $\mathcal{U}$. We then resample $\mathbf{s}$ from the initial $\mathcal{U}$ to create a final $\mathcal{U}$ such that the predicted labels of $A(\mathbf{s})$ has the marginal distribution $\mathbf{q}_{\mathcal{U}}$. Next, from Bayes theorem,

$$p(G(\mathbf{s}) = i \mid A(\mathbf{s}) = j) = \frac{m_{ij}p_{\mathcal{U},i}}{\sum_i m_{ij}p_{\mathcal{U},i}} = \alpha_{ij}.$$

For $\mathbf{s}$ in $\mathcal{U}$ such that $A(\mathbf{s}) = j$, we generate the new "true" labels from Multinomial$(1, (\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{Cj})')$. This data generation process ensures that for any $\mathbf{s}$ in $\mathcal{U}$ both $G(\mathbf{s}) \sim$ Multinomial$(1, \mathbf{p}_{\mathcal{U}})$ and $A(\mathbf{s}) \mid G(\mathbf{s}) = i \sim$ Multinomial$(1, \mathbf{M}_{i*})$ are approximately true. We repeat the procedure for $\mathcal{L}$, using the same $\mathbf{M}$ but a different $\mathbf{p}_{\mathcal{L}}$. This reflects the reality for verbal autopsy data where the symptom-given-cause dynamics is same for all deaths $\mathcal{L} \cup \mathcal{U}$ in the new country, but the hospital distribution of causes $\mathbf{p}_{\mathcal{L}}$ is unlikely to match the population CSMF $\mathbf{p}_{\mathcal{U}}$. For resampling to create the final $\mathcal{L}$, we also vary n — the size of $\mathcal{L}$ as 50, 100, 200 and 400, to represent varying amount of local labeled that will be available at different stages of a project.

We consider two choices of A: Tariff (version 1.0.3) and InSilicoVA (version 1.2.2). For $\mathbf{M}$, we use three choices. We have $\mathbf{M}_1 = \mathbf{I}$,

$$\mathbf{M}_2 = \begin{pmatrix} 1.00 & 0 & 0 & 0 \\ 0.65 & 0.35 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

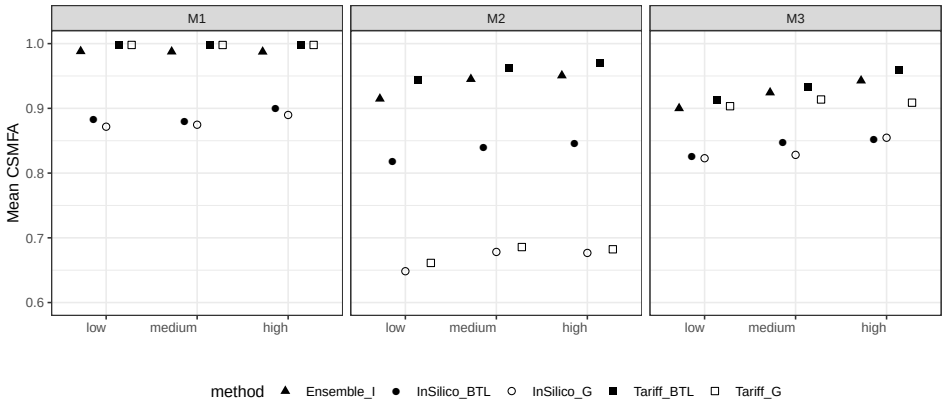
and $\mathbf{M}_3 = 0.6 * \mathbf{I} + 0.1 * \mathbf{1}\mathbf{1}'$. The first choice represents the case where the algorithm A is perfect for predicting in the target domain. $\mathbf{M}_2$ with

two large off-diagonal entries and all other off-diagonal ones being zero represents the scenario where there are one or two systematic sources of bias in A when trained on a source domain $\mathcal{D}_S$ different from $\mathcal{D}_T$. The specific choice of $\mathbf{M}_2$ depicts the scenario that 65% of Diarrhea/Dysentry cases are classified as Pneumonia and 50% of Sepsis deaths are categorized as some other cause. Finally, $\mathbf{M}_3$ represents the scenario where there are many small misclassifications.

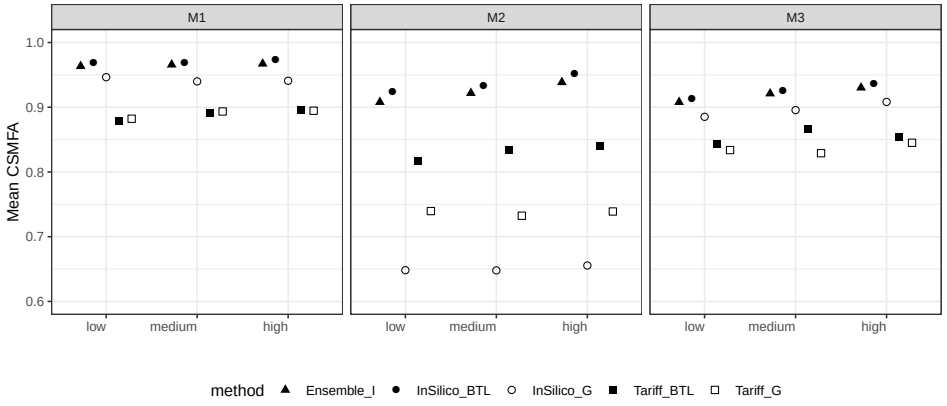
To ensure that $\mathbf{p}_\mathcal{U}$ and $\mathbf{p}_\mathcal{L}$ are different, we generate pairs of probability vectors $(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{U})'$ from Dirichlet($\mathbf{1}$) distribution and divide the cases into three scenarios: *low:* CSMFA$(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{U}) < 0.4$, *medium:* $0.4 <$ CSMFA$(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{U}) < 0.6$, and *high:* CSMFA$(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{U}) > 0.6$. Here CSMFA denoting the CSMF accuracy is a metric quantifying the distance of a probability vector ($\mathbf{p}_\mathcal{L}$) from a reference probability vector ($\mathbf{p}_\mathcal{U}$) and is given by (Murray et al., 2011a):

$$CSMFA(\mathbf{p}_\mathcal{L}, \mathbf{p}_\mathcal{U}) = 1 - \frac{||\mathbf{p}_\mathcal{L} - \mathbf{p}_\mathcal{U}||_1}{2(1 - \min \mathbf{p}_\mathcal{U})}.$$

For each scenario, we generated 100 pairs of $\mathbf{p}_\mathcal{L}$ and $\mathbf{p}_\mathcal{U}$. For each generated dataset, we use all the algorithms listed in Table 2.2 for predicting $\mathbf{p}_\mathcal{U}$. For an estimate $\widehat{\mathbf{p}}_\mathcal{U}(x)$ generate by a model x, we assess the performance of x using CSMFA(x)= CSMFA$(\widehat{\mathbf{p}}_\mathcal{U}(x), \mathbf{p}_\mathcal{U})$. We present a brief summary of the results here. A much more detailed analysis is provided in Section 2.9.5 of the supplementary material. Figure 2.2 presents the CSMFA for all the five models for $n = 400$. The three columns are for the three choices of $\mathbf{M}$ described above, and in each figure the x-axis from left to right marks the *low, medium* and *high* settings.



(a) Data generated using Tariff



(b) Data generated using InSilicoVA

Figure 2.2: CSMF of ensemble and single-classifier transfer learners.

We observe that for almost all settings the Bayesian transfer learning approach was better than its corresponding baseline, i.e. Tariff$_{BTL}$ was better than Tariff$_{\mathcal{G}}$ and InSilicoVA$_{BTL}$ was better than InSilicoVA$_{\mathcal{G}}$. The improvement in CSMFA was most drastic for $\mathbf{M}_2$ (middle column) where it was as much as 0.3 in some cases. Only for $\mathbf{M}_1$, i.e., when the classifier is assumed to be perfect for predicting in the target domain, we see Tariff$_{BTL}$ and Tariff$_{\mathcal{G}}$ produce similar CSMFA in the (top-left) and InSilicoVA$_{BTL}$ and InSilicoVA$_{\mathcal{G}}$ produce similar CSMFA (bottom-left). This just corroborates Theorem 1 that the transfer learning keeps things unchanged if the classifier has zero transfer error. We also observe that within each figure, CSMFA's generally increase as we go from the low to the high setting, indicating that increased representativeness of the class distribution in the small labeled set $\mathcal{L}$ leads to improved performance. Also, across all settings we see that transfer learning based on algorithms used to simulate the data performs better, i.e., for the top-row Tariff$_{BTL}$ performs better than InSilicoVA$_{BTL}$ as in this case they respectively correspond to a true and a misspecified model. Similarly, for the bottom-row InSilicoVA$_{BTL}$ performs better than Tariff$_{BTL}$. However, even under model misspecification, the transfer learning models perform better than their baselines, i.e., even when data is generated using Tariff, InSilicoVA$_{BTL}$ performs better than InSilicoVA$_{\mathcal{G}}$. Finally, across all scenarios, the ensemble learner performs close to the better performing individual learner, highlighting its utility and robustness.

In Section 2.9.5 of the supplementary material (http://www.biostatistics.oxfordjournals.org) we present more thorough insights into the simulation study. In Section 2.9.5.1

we assess the impact of the disparity in the class distributions between the source and target domains. In Section 2.9.5.2 we compare the biases in the estimates of individual class probabilities. Section 2.9.5.3 delves into the role of the sample size and quality of the limited labeled set $\mathcal{L}$. Section 2.9.5.4 demonstrates the value of the Bayesian shrinkage by comparing with the frequentist transfer learning outlined in Section 2.2.1. In Section 2.9.5.5 we compare the joint and independent ensemble models and demonstrate how they favorably weight the more accurate algorithm. Section 2.9.5.6 shows how one can use informed shrinkage, if a practitioner has apriori knowledge of which causes are likely to be misclassified by an algorithm. Finally, in Section 2.9.5.7, we compare the performance of the models for predicting individual-level class probabilities for target domain data using the algortihm outlined in Section 2.9.2.

2.6 Predicting CSMF in India and Tanzania

We evaluate the performance of baseline CCVA algorithms and our transfer learning approach when predicting the CSMF for under 5 children in India and Tanzania using the PHMRC data with actual COD labels. We used both India and Tanzania, as they were the only countries with substantial enough sample sizes ($N_{India} = 948$, $N_{Tanzania} = 728$). For a given country (either India or Tanzania), we first split the PHMRC child data into subjects from within the country ($\mathcal{L}$ and $\mathcal{U}$) and from outside of the country ($\mathcal{G}$). We then used weighted sampling to select $n(\in \{50, 100, 200\})$ subjects from within the country of interest to be in $\mathcal{L}$, using weights such that $\text{CSMFA}(\mathbf{p}_{\mathcal{L}}, \mathbf{p}_{\mathcal{U}})$ was low. Figure 2.11

in the supplementary material (http://www.biostatistics.oxfordjournals.org) shows the difference in the marginal symptom distribution between $\mathcal{U}$ and $\mathcal{L}$. All the subjects from the country were put in $\mathcal{U}$. We trained models $Insilico_{\mathcal{G}}$ and $Tariff_{\mathcal{G}}$ using the non local data $\mathcal{G}$, which were then used to predict the top COD for all subjects in $\mathcal{L}$ and $\mathcal{U}$. We classified all causes of death into "External", "Pneumonia", "Diarrhea/Dysentery", "Other Infectious", and "Other". These predictions were then used to estimate the baseline CSMFs and as an input to our transfer learning models Tariff$_{BTL}$, InSilicoVA$_{BTL}$, and $Ensemble_I$. Since the true labels (GS-COD) are available in PHMRC, we calculated the true $\mathbf{p}_{\mathcal{U}}$ for a country as the empirical proportions of deaths from each cause, based on all the records within the country. This $\mathbf{p}_{\mathcal{U}}$ was used to calculate the CSMF accuracy of each model. This whole process was repeated 500 times for each combination of country and value of n. This made sure that the results presented are average over 500 different random samples of $\mathcal{L}$ for each country, and are not for an arbitrary sample.

Figure 2.3 presents the results of this analysis. The top and bottom rows represent the results for India and Tanzania respectively. The four columns correspond to four different choices of n. There are several notable observations. First, regardless of n, choice of algorithm A, and country, the calibrated estimates of prevalence from our transfer learning model performed better than or similar to the analogous baseline CSMFs, i.e., Tariff$_{BTL}$ performed better than Tariff$_{\mathcal{G}}$, and InSilicoVA$_{BTL}$ performed better than InSilicoVA$_{\mathcal{G}}$. Second, the magnitude of improvement for the our approach depends on the country and the size of $\mathcal{L}$. Within India, the CSMFA of Tariff$_{BTL}$ and InSilicoVA$_{BTL}$ is

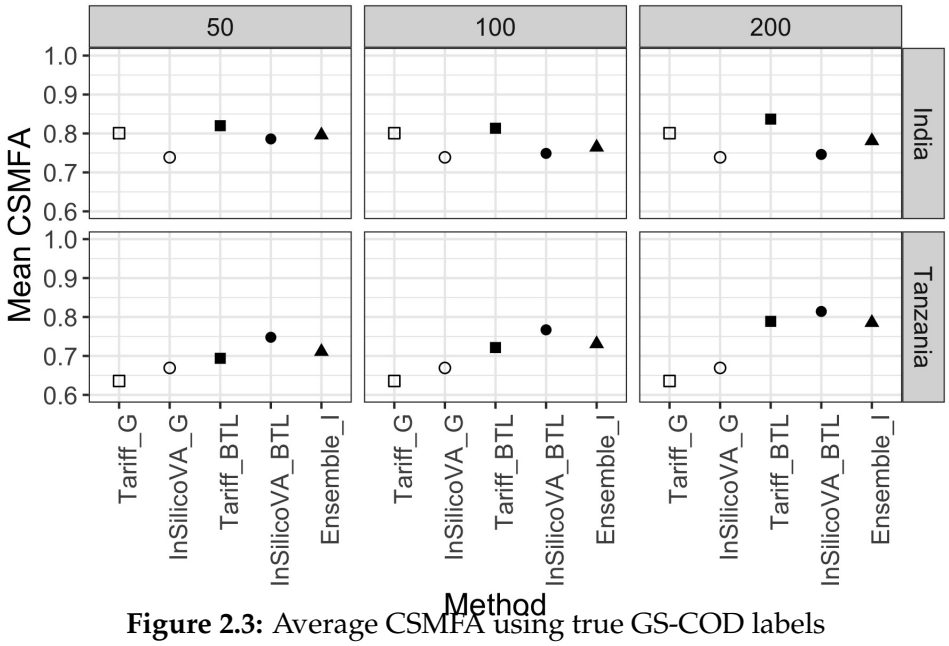


Figure 2.3: Average CSMFA using true GS-COD labels

similar to respectively those from Tariff$_{\mathcal{G}}$ and InSilicoVA$_{\mathcal{G}}$. Tariff does better than InsilicoVA for India with Tariff$_{BTL}$ being the best performer. In Tanzania, the baseline InsilicoVA model InSilicoVA$_{\mathcal{G}}$ does better than Tariff$_{\mathcal{G}}$, and similarly InSilicoVA$_{BTL}$ does better than InSilicoVA$_{BTL}$. The improvement of Tariff$_{BTL}$ and InSilicoVA$_{BTL}$ respectively over Tariff$_{\mathcal{G}}$ and InSilicoVA$_{\mathcal{G}}$ is more prominent than in India, with InSilicoVA$_{BTL}$ being the most accurate. The magnitude of improvement in the three TL approaches also increased with increase in n for Tanzania.

2.7 Discussion

Epidemiological studies pose unique challenges to transfer learning, stemming from its focus on estimating population-level quantities as opposed to individual predictions, small sample sizes coupled with high-dimensional

covariate spaces (survey records), and lack of large training databases available for many other machine learning tasks. Motivated by these settings, we have presented a parsimonious hierarchical model-based approach to transfer learning of population-level class probabilities, using a pre-trained classifier, limited labeled target domain data, and abundant unlabeled target domain data.

In order for the transfer learning approach to work, the labeled data $\mathcal{L}$ has to be useful for improving CSMF estimation in $\mathcal{U}$, i.e., there has to be some commonality between the distributions in $\mathcal{L}$ and $\mathcal{U}$. Usually $\mathcal{L}$ is never going to be representative of the marginal cause distribution of $\mathcal{U}$. If additionally, it is also not representative of the conditional distributions of $\mathbf{s} \mid G(\mathbf{s})$ (or, in our dimension reduction approach, of $A(\mathbf{s}) \mid G(\mathbf{s})$), then $\mathcal{L}$ is of no use to improve CSMF estimation in $\mathcal{U}$. Hence, our transfer learning is useful when the the conditional distributions are same (constant $\mathbf{M}$) between $\mathcal{L}$ and $\mathcal{U}$, or has the same functional form (regression approach of Section 2.4.2) between $\mathcal{L}$ and $\mathcal{U}$.

Shrinkage or regularization is at the core of our approach. In datasets with large numbers of variables (dimensions), regularized methods have become ubiquitous. A vast majority of the literature focuses on shrinking estimates (mostly regression coefficients and covariance or precision matrices) towards some known sub-model. We apply the same principle of regularization in a unique way for estimating the population class probabilities. Instead of shrinking towards any underlying assumptions about the true population distribution, we shrink towards the baseline estimate from a classifier trained on source data. In absence of sufficient target-domain data, this is the best

available estimate and has to be used. We show how this shrinkage for the class probabilities is equivalent to shrinking the confusion matrix towards the identity matrix and construct appropriate Dirichlet priors to achieve this shrinkage. This regularized estimation of a confusion matrix (or any transition matrix) can also be applied in other contexts.

The fully Bayesian implementation is fast, owing to a novel data-augmented Gibbs sampler. The ensemble model ensures robust estimates via data-driven averaging over many classifiers and reduces the risk of selecting a poor one for a particular application. Our simulations demonstrate the value of transfer learning, offering substantially improved accuracy. The PHMRC data analysis makes evident the value of collecting a limited number of labeled data GS-COD in the local population using full or minimally invasive autopsies, alongside the nationwide VA survey. Subsequently using transfer learning improves the CSMF estimates. The results also show how our approach benefits from larger sample sizes of the local labeled set $\mathcal{L}$, and from closer alignment between the marginal class probabilities in $\mathcal{L}$ and the true target domain class probabilities.

For VA data, we note that while we have treated $\mathcal{G}$ as a labeled dataset in the source domain $\mathcal{D}_S$, in practice it can be any other form of gold standard information sufficient to train a VA classifier. CCVA methods like Tariff and the approach in King, Lu, et al., 2008 represent a traditional supervised learning approach and needs a substantial labeled training dataset $\mathcal{G}$. InterVA is a semi-supervised learning approach where $\mathcal{G}$ is a standard matrix of letter grades representing the propensity of each symptom given each cause. InSilicoVA

generalizes InterVA and endows the problem with a proper probabilistic framework allowing coherent statistical inference. It adapts to the type of $\mathcal{G}$ and can work with either the default symptom-cause matrix used in InterVA or estimate this matrix based on some labeled training data of paired VA and GS-COD records. Our transfer learning is completely agnostic to the choice of this baseline CCVA algorithm and the form of $\mathcal{G}$ they require. We only need the predictions from a pre-trained algorithm for all observations in $\mathcal{L} \cup \mathcal{U}$.

One important direction forward would be to generalize this approach for more complex COD outcomes. Currently COD outcome is viewed as a discrete variable taking values on a set of causes like Pneumonia or Sepsis. In practice, death is a complex chronological series of several events starting from some root causes and ending at the immediate or proximal cause. In addition to understanding prevalence of causes in the population, another goal for many of the aforementioned programs is to identify medical events that occurred before death for which an intervention could prevent or delay mortality. Extending the current setup for hierarchical or tree-structured COD outcome would be a useful tool to address this aim. Many CCVA algorithms, in addition to predicting the most likely COD, also predict the (posterior) distribution of likely causes. Our current implementation only uses the most likely COD as an input. An extension enabling the use of the full predictive distribution as an input can improve the method. In particular, this will benefit the individual COD predictions for which currently two individuals with the same predicted COD from CCVA have the same predicted COD distribution after transfer learning. Finally, the VA records, containing about

250 questions for thousands of individuals, naturally has several erroneous entries. Currently preprocessing VA records to eliminate absurd entries and records entails onerous manual labor. It is challenging to develop quality control models for VA data due to the high dimensionality of the symptoms. Akin to what we did here, one can consider dimension reduction via the predictions of CCVA algorithms for an automated statistical quality control for VA records.

2.8 Software

R-package 'calibratedVA' containing code to obtain estimates of population CSMFs from our transfer learning approach using baseline predictions from any verbal autopsy algorithm is available at https://github.com/jfiksel/CalibratedVA/. The package also contains the code for the ensemble model for using outputs from several VA algorithms. A vignette describing how to navigate the package and demonstrating the use of the methodology is provided in https://github.com/jfiksel/CalibratedVA/blob/master/vignettes/CalibratedVA.Rmd. All results in this paper can be recreated using the scripts contained in https://github.com/jfiksel/BayesTLScripts.

2.9 Supplementary Material

2.9.1 MAP estimation

In the main manuscript, we have only discussed fully Bayesian implementations of the model in (2.4). If full inferential output is superfluous and

only posterior point-estimates of the parameters are desired, we outline a MAP (Maximum a posteriori) estimation for obtaining posterior modes of the parameters using an EM-algorithm. The data augmentation scheme introduced for the Gibbs sampler in 2.2.3 is also seamlessly congruous with the EM algorithm.

In particular, we consider the vector $\mathbf{v}$ and $\mathbf{T}$ as the observed data and augment $\mathbf{B}$ introduced in Section 2.2.3 as the missing data to form the complete data likelihood $l(\mathbf{B}, \mathbf{v}, \mathbf{T} \mid \mathbf{M}, \mathbf{p}, \boldsymbol{\gamma})$ which is proportional to (2.5). At the s^{th} iteration, let $\mathbf{M}^{[s]} = (m_{ij}^{[s]})$, $\mathbf{p}^{[s]} = (p_i^{[s]})$ denote the current values of the parameters. Then

$$E^{[s]}(b_{ij} \mid \mathbf{v}, \mathbf{T}) = \frac{v_j m_{ij}^{[s]} p_i^{[s]}}{\sum_i m_{ij}^{[s]} p_i^{[s]}} = \widehat{b}_{ij}^{[s]},$$

where $E^{[s]}$ denotes the expectation taken using the parameter values from the s^{th} iteration. The EM algorithm then proceeds as follows:

$$\begin{aligned} \text{E-step: } E^{[s]}(\log l(\mathbf{B}, \mathbf{v}, \mathbf{T} \mid \mathbf{M}, \mathbf{p}, \boldsymbol{\gamma}) \mid \mathbf{v}, \mathbf{T}) = \sum_i \Bigg(\sum_j \Big(\widehat{b}_{ij}^{[s]} \log(m_{ij} p_i) + \\ (t_{ij} + \gamma_i \epsilon + \gamma_i I(i=j) - 1) \log(m_{ij}) \Big) + (\delta - 1) \log p_i + h(\gamma_i) \Bigg) \end{aligned} \tag{2.13}$$

where $h(\gamma) = \log\left(\frac{\Gamma(C\gamma+\epsilon)}{\Gamma(\gamma\epsilon)^{C-1}\Gamma(\gamma\epsilon+\gamma)}\right) + (\alpha - 1)\log\gamma - \beta\gamma$. Subsequently, the

maximization step can be formulated as:

$$\text{M-step:} \quad \begin{aligned} m_{ij}^{[s+1]} &= \frac{\widehat{b}_{ij}^{[s]} + t_{ij} + \gamma_i \epsilon + \gamma_i I(i=j) - 1}{\sum_j (\widehat{b}_{ij}^{[s]} + t_{ij}) + \gamma_i C\epsilon + \gamma_i - C} \\ p_i^{[s+1]} &= \frac{\sum_j \widehat{b}_{ij}^{[s]} + \delta - 1}{\sum_i \sum_j \widehat{b}_{ij}^{[s]} + C\delta - C} \\ \gamma_i^{[s+1]} &= \arg\max_\gamma \textstyle\sum_j (\gamma\epsilon + \gamma I(i=j) - 1) \log(m_{ij}) + h(\gamma) \end{aligned} \tag{2.14}$$

The closed form expression of $\mathbf{M}$ and $\mathbf{p}$ in the M-step is a consequence of the data augmentation. This drastically accelerates the MAP estimation as we only need to conduct C univariate optimizations, one corresponding to each γ_i. If instead the data augmentation was not exploited and only the observed likelihood was used, we would need to search an $O(C^2)$ dimensional space to obtain the MAP estimates. We can implement similar MAP estimation algorithms for the joint and independent ensemble models detailed in Section 2.3. We omit the steps here.

2.9.2 Individual-level transfer learning

While our Bayesian transfer learning is primarily targeted to estimate population-level class probabilities, it can also be used to predict individual-level class probabilities in the target domain $\mathcal{D}_T$. The posterior distribution of the true class membership $G(\mathbf{s}_r)$ of the r^{th} individual is given by

$$\begin{aligned} p(G(\mathbf{s}_r) = i \mid A(\mathbf{s}_r) = j, \mathbf{T}) &= \int p(G(\mathbf{s}_r) = i \mid \mathbf{p}, \mathbf{M}, \boldsymbol{\gamma}, A(\mathbf{s}_r) = j, \mathbf{T}) \times \\ &\qquad p(\mathbf{p}, \mathbf{M}, \boldsymbol{\gamma} \mid \mathbf{v}, \mathbf{T})\, dP(\mathbf{M})\, dP(\mathbf{p})\, dP(\boldsymbol{\gamma}) \\ &= \int p(G(\mathbf{s}_r) = i \mid \mathbf{p}, \mathbf{M}, A(\mathbf{s}_r) = j) p(\mathbf{p}, \mathbf{M} \mid \mathbf{v}, \mathbf{T})\, dP(\mathbf{M})\, dP(\mathbf{p}) \\ &= \int \frac{m_{ij} p_i}{\sum_{i=1}^{C} m_{ij} p_i} p(\mathbf{p}, \mathbf{M} \mid \mathbf{v}, \mathbf{T})\, dP(\mathbf{M})\, dP(\mathbf{p}) \end{aligned}$$

We can now easily conduct composition sampling using posterior samples of $\mathbf{M}$ and $\mathbf{p}$ to generate a posterior distribution for $G(\mathbf{s}_r)$. This simple application of the Bayes theorem, can recover the individual class memberships. However, it is a crude approach because the posterior distribution of the $G(\mathbf{s}_r)$ are identical for all instances $\mathbf{s}_r$ with the same predicted class $A(\mathbf{s}_r)$ from A. If A is a probabilistic classifier like InSilicoVA (McCormick et al., 2016), then in addition to providing a predicted class membership $A(\mathbf{s}_r)$, A also provides the predicted distribution for each individual's class. Utilizing the entire predicted distribution from A should lead to improved individual level transfer learning. Since the focus of this manuscript is population level transfer learning, we do not further explore this avenue here.

2.9.3 Gibbs sampler for the joint ensemble model

Let $y_{\mathbf{j}}$ be the number of instances in $\mathcal{U}$ for which algorithm $A^{(1)}$ predicts cause j_1, $A^{(2)}$ predicts cause j_2, and so on. Let $\mathbf{y}^*$ be the $C^K \times 1$ vector formed by stacking the $y_{\mathbf{j}}$'s. Also, let $u_{i\mathbf{j}} = \prod_{k=1}^{K} m_{ij_k}^{(k)}$ and $\mathbf{u}_{\mathbf{j}} = (u_{1\mathbf{j}}, u_{2\mathbf{j}}, \ldots, u_{C\mathbf{j}})'$.

The posterior $\mathbf{p}, \{\mathbf{M}^{(k)}, \boldsymbol{\gamma}_k\}_{k=1,\ldots,K} \mid \mathbf{T}^{(1)}, \mathbf{T}^{(2)}, \ldots, \mathbf{T}^{(k)}, \mathbf{w}^*$ is proportional to

$$\prod_{\mathbf{j}} (\sum_i u_{i\mathbf{j}} p_i)^{y_{\mathbf{j}}} \times \prod_i p_i^{\delta-1} \times$$

$$\prod_{k=1}^{K} \left(\prod_{i=1}^{C} \frac{\Gamma(\gamma_i^{(k)}(C\epsilon+1))}{(\Gamma(\gamma_i^{(k)}\epsilon))^{C-1}\Gamma(\gamma_i^{(k)}(\epsilon+1))} \times \prod_j (m_{ij}^{(k)})^{t_{ij}^{(k)}+\gamma_i^{(k)}(\epsilon+1(i=j))-1} \right).$$

We will once again use data augmentation to implement the Gibbs sampler.

Let $\mathbf{b_j} = (b_{1\mathbf{j}}, b_{2\mathbf{j}}, \ldots, b_{C\mathbf{j}})')$ denote the $C \times 1$ dimensional realization of a Multinomial $(y_\mathbf{j}, \mathbf{1}/C)$ distribution, and let $\mathbf{B}$ denote the $C^K \times C$ matrix formed by stacking the independent $\mathbf{b_j}$'s row-wise for all combinations of $\mathbf{j}$. Then we have the following full conditionals for the Gibbs sampler:

$$\mathbf{b_j} \mid \cdot \sim Multinomial(y_\mathbf{j}, \frac{1}{\mathbf{1}'(\mathbf{u_j} \odot \mathbf{p})}\mathbf{u_j} \odot \mathbf{p})$$

$$\mathbf{M}_{i*}^{(k)} \mid \cdot \sim Dirichlet\left(\mathbf{T}_{i*} + \gamma_i^{(k)}\mathbf{I}_{i*} + \gamma_i^{(k)}\mathbf{1} + (\sum_{\mathbf{j}:j_k=1} b_{i\mathbf{j}}, \ldots, \sum_{\mathbf{j}:j_k=C} b_{i\mathbf{j}})'\right)$$

$$\mathbf{p} \mid \cdot \sim Dirichlet(\sum_\mathbf{j} b_{1\mathbf{j}} + \delta, \ldots, \sum_\mathbf{j} b_{C\mathbf{j}} + \delta)$$

Here $\odot$ denotes the Hadamard (elementwise) product.

Finally, as in Section 2.2.2, we update $\gamma_i^{(k)}$'s using a metropolis random walk with log-normal proposal to sample from the full conditionals

$$p(\gamma_i^{(k)} \mid \cdot) \propto \frac{\Gamma(C\gamma_i^{(k)}\epsilon + \gamma_i)}{\Gamma(\gamma_i^{(k)}\epsilon)^{C-1}\Gamma(\gamma_i^{(k)}\epsilon + \gamma_i^{(k)})} \times$$
$$(\gamma_i^{(k)})^{\alpha-1} \exp(-\beta\gamma_i^{(k)}) \prod_j (m_{ij}^{(k)})^{\gamma_i^{(k)}\epsilon + \gamma_i^{(k)} 1(i=j)}.$$

2.9.3.1 Individual level classifications

As illustrated in Section 2.9.2, the ensemble transfer learner can also predict the individual-level class memberships. Using Bayes theorem we have

$$p(G(\mathbf{s}_r) = i \mid a_r^{(1)} = j_1, \ldots, a_r^{(K)} = j_k) = \frac{1}{\sum_\mathbf{j} u_{i\mathbf{j}} p_i} u_{i\mathbf{j}} p_i.$$

Since posterior distributions of $u_{i\mathbf{j}}$'s and $\mathbf{p}$ have already been sampled, we can generate posterior samples of $G(\mathbf{s}_r)$ post-hoc using the composition sampling approach demonstrated in Section 2.9.2.

For the independent ensemble model, one can recover the posterior distribution of the individual class memberships in the exact same way. Only additional step would be to first calculate the $\mathbf{u_j}$'s as they are no longer part of the Gibbs sampler.

2.9.4 Proofs

Theorem 1. The marginal posterior $\mathbf{p} \mid \mathbf{v}, \mathbf{T}$ is given by $\int p(\mathbf{p}, \mathbf{M}, \boldsymbol{\gamma} \mid \mathbf{v}, \mathbf{T}) dP(\mathbf{M}) dP(\boldsymbol{\gamma})$. Conditional on $\boldsymbol{\gamma}$, looking only at terms that involve $\mathbf{p}$, $\mathbf{M}$, and $\boldsymbol{\gamma}$, we have

$$p(\mathbf{p}, \mathbf{M}, \mathbf{v}, \mathbf{T} \mid \boldsymbol{\gamma}) \propto \prod_j (\sum_i m_{ij} p_i)^{v_j} \times \prod_i p_i^{\delta - 1} \times$$

$$\prod_i \frac{\Gamma(\gamma_i(C\epsilon + 1))}{(\Gamma(\gamma_i \epsilon))^{C-1} \Gamma(\gamma_i(\epsilon + 1))} \prod_j (m_{ij})^{t_{ij} + \gamma_i(\epsilon + 1(i=j)) - 1}$$

We will now use the multinomial theorem to expand the first product $\prod_j (\sum_i m_{ij} p_i)^{v_j}$. Note that the j^{th} term expands into $\kappa_j = \binom{v_j + C - 1}{C - 1}$ terms, one corresponding to each partition of v_j. Let $\mathbf{B}^{(j)} = (b_{ki}^{(j)})$ denote the $\kappa_j \times C$ partition matrix formed by stacking up all $1 \times C$ rows that represent a non-negative integer partition of v_j. The k^{th} row of $\mathbf{B}^{(j)}$ gives the k^{th} partition and i^{th} element of that row corresponds to power index for the i^{th} term $(m_{ij} p_i)$. We now have,

$$p(\mathbf{p}, \mathbf{M} \mid \mathbf{v}, \mathbf{T}, \boldsymbol{\gamma}) \propto \left(\prod_j \sum_{k_j=1}^{\kappa_j} \prod_i \frac{(m_{ij} p_i)^{b_{k_j i}^{(j)}}}{b_{k_j i}^{(j)}!} \right) \times \prod_i p_i^{\delta-1} \times$$

$$\prod_i \frac{\Gamma(\gamma_i(C\epsilon+1))}{(\Gamma(\gamma_i\epsilon))^{C-1}\Gamma(\gamma_i(\epsilon+1))} \prod_j (m_{ij})^{t_{ij}+\gamma_i(\epsilon+1(i=j))-1}$$

$$\propto \sum_{k_1=1}^{n_1} \cdots \sum_{k_C=1}^{n_C} \left(\prod_i \frac{\Gamma(\gamma_i(C\epsilon+1)) p_i^{\sum_j b_{k_j i}^{(j)}-1}}{(\Gamma(\gamma_i\epsilon))^{C-1}\Gamma(\gamma_i(\epsilon+1))} \times \right.$$

$$\left. \prod_j \frac{(m_{ij})^{b_{k_j i}^{(j)}+t_{ij}+\gamma_i(\epsilon+1(i=j))-1}}{b_{k_j i}^{(j)}!} \right)$$

Given $k_1, \ldots, k_C$ and i, the product $\prod_{j=1}^{C} (m_{ij})^{b_{k_j i}^{(j)}+t_{ij}+\gamma_i(\epsilon+1(i=j))-1}$ is the kernel of a $Dirichlet(b_{k_1 i}^{(1)} + t_{i1} + \gamma_i\epsilon, \ldots, b_{k_i i}^{(i)} + t_{ii} + \gamma_i(\epsilon+1), \ldots, b_{k_C i}^{(C)} + t_{iC} + \gamma_i\epsilon)$ distribution. Hence, integrating $\mathbf{M}$ out with respect to the order $\prod_{i=1}^{C} \prod_{j=1}^{C} dm_{ij}$, we are left with

$$p(\mathbf{p} \mid \mathbf{v}, \mathbf{T}, \boldsymbol{\gamma}) \propto \sum_{k_1=1}^{n_1} \cdots \sum_{k_C=1}^{n_C} w_{k_1,k_2,\ldots,k_C}(\boldsymbol{\gamma}, \epsilon) \prod_i p_i^{\sum_j b_{k_j i}^{(j)}+\delta-1}$$

where $w_{k_1,k_2,\ldots,k_C}(\boldsymbol{\gamma},\epsilon) = \prod_i \frac{\Gamma(\gamma_i(C\epsilon+1))\prod_{j=1}^{C}\Gamma(b_{k_ji}^{(j)}+t_{ij}+\gamma_i(\epsilon+1(i=j)))}{(\Gamma(\gamma_i\epsilon))^{C-1}\Gamma(\gamma_i(\epsilon+1))\Gamma(\sum_j(b_{k_ji}^{(j)}+t_{ij})+\gamma_i(C\epsilon+1))\prod_j b_{k_ji}^{(j)}!}$. Hence,

$$\mathbf{p} \mid \mathbf{v},\mathbf{T} \sim \sum_{k_1=1}^{n_1}\cdots\sum_{k_C=1}^{n_C}\left(\left(\int \frac{1}{W(\boldsymbol{\gamma},\epsilon)} w_{k_1,k_2,\ldots,k_C}(\boldsymbol{\gamma},\epsilon)dF(\boldsymbol{\gamma})\right)\times \right.$$

$$\left. Dirichlet(\sum_j b_{k_j1}^{(j)}+\delta,\ldots,\sum_j b_{k_jC}^{(j)}+\delta)\right)$$

where $W(\boldsymbol{\gamma},\epsilon) = \sum_{k_1=1}^{n_1}\cdots\sum_{k_C=1}^{n_C} w_{k_1,k_2,\ldots,k_C}(\boldsymbol{\gamma},\epsilon)$. Without loss of generality, let the first row of each $\mathbf{B}^{(j)}$ represent the partition of v_j which allocates v_j to the j^{th} component and 0 to all the other components. For any $(k_1,k_2,\ldots,k_C)' \neq \mathbf{1}_C$, we have

$$\lim_{\epsilon\to 0}\frac{w_{k_1,k_2,\ldots,k_C}(\boldsymbol{\gamma},\epsilon)}{w_{1,1,\ldots,1}(\boldsymbol{\gamma},\epsilon)} = \prod_i\left(\frac{\Gamma(\sum_j t_{ij}+v_i+\gamma_i)\Gamma(b_{k_ii}^{(i)}+t_{ii}+\gamma_i)}{\Gamma(\sum_j(b_{k_ji}^{(j)}+t_{ij})+\gamma_i)\Gamma(v_i+t_{ii}+\gamma_i)}\times\right.$$

$$\left.\left(\prod_{j\neq i}\lim_{\epsilon\to 0}\frac{\Gamma(b_{k_ji}^{(j)}+t_{ij}+\gamma_i\epsilon)}{\Gamma(t_{ij}+\gamma_i\epsilon)}\right)\right)$$

If $b_{k_ji}^{(j)} = 0$, the ratio $\frac{\Gamma(b_{k_ji}^{(j)}+t_{ij}+\gamma_i\epsilon)}{\Gamma(t_{ij}+\gamma_i\epsilon)}$ is one. However, since $(k_1,k_2,\ldots,k_C)' \neq \mathbf{1}_C$, we have atleast one pair $i \neq j$ such that $b_{k_ji}^{(j)} \geq 1$ and consequently

$$\frac{\Gamma(b_{k_ji}^{(j)}+t_{ij}+\gamma_i\epsilon)}{\Gamma(t_{ij}+\gamma_i\epsilon)} = \prod_{s=0}^{b_{k_ji}^{(j)}-1}(s+t_{ij}+\gamma_i\epsilon) \xrightarrow{\epsilon\to 0} 0$$

since T is diagonal. Hence, $w_{1,1,\ldots,1}$ dominates all the other weights in the limiting case. Since each of the scaled weights are less than one, using dominated

convergence theorem,

$$\lim_{\epsilon \to 0} \int \frac{1}{W(\gamma, \epsilon)} w_{k_1, k_2, \ldots, k_C}(\gamma, \epsilon) dF(\gamma) = 1((k_1, k_2, \ldots, k_C)' = \mathbf{1})$$

and hence $\lim_{\epsilon \to 0} p(\mathbf{p} \mid \mathbf{v}, \mathbf{T}) \propto \prod_i p_i^{\sum_j b_{1i}^{(j)} + \delta - 1} = \prod_i p_i^{v_i + \delta - 1}$. □

Theorem 2. We proof only for the case $K = 2$ as the same proof generalizes for arbitrary K. We simplify the notation for the proof. Let v_{st} denote the number of instances in $\mathcal{U}$ assigned to class s by algorithm 1, and class t by algorithm 2. We write $\mathbf{M}^{(1)} = \mathbf{M}$, $\mathbf{M}^{(2)} = \mathbf{N}$, $\mathbf{T}^{(1)} = \mathbf{T}$ and $\mathbf{T}^{(2)} = \mathbf{U}$ to get rid of the superscripts. Also, let $\mathbf{B}_{(st)} = (b_{li}^{(st)})$ denote a $\kappa_{st} \times C$ matrix formed by stacking row-wise all possible partitions of v_{st} into C non-negative integers. Here $\kappa_{st} = \binom{v_{st}+C-1}{C-1}$ denotes the total number of such partitions. Let $\mathbf{h} = (h_{11}, h_{12}, \ldots, h_{CC})'$ denote a generic index vector such that each $h_{st} \in \{1, 2, \ldots, \kappa_{st}\}$ indexes a partition of v_{st} and $\mathcal{H}$ denote the collection of all such $\mathbf{h}$'s. Then likelihood for $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N)'$ is

$$\begin{aligned}
\prod_{s=1}^{C} \prod_{t=1}^{C} \left(\sum_{l=1}^{\kappa_{st}} \prod_{i=1}^{C} \frac{(p_i m_{is} n_{it})^{b_{li}^{(st)}}}{b_{li}^{(st)}!} \right) &= \sum_{\mathbf{h} \in \mathcal{H}} \prod_{i=1}^{C} \prod_{s=1}^{C} \prod_{t=1}^{C} \frac{(p_i m_{is} n_{it})^{b_{h_{st} i}^{(st)}}}{b_{h_{st} i}^{(st)}!} \\
&= \sum_{\mathbf{h} \in \mathcal{H}} \left(\frac{1}{c_{\mathbf{h}}} \prod_{i=1}^{C} p_i^{\sum_{s=1}^{C} \sum_{t=1}^{C} b_{h_{st} i}^{(st)}} \times \right. \\
&\quad \left. \prod_{s=1}^{C} m_{is}^{\sum_{t=1}^{C} b_{h_{st} i}^{(st)}} \prod_{t=1}^{C} n_{it}^{\sum_{s=1}^{C} b_{h_{st} i}^{(st)}} \right)
\end{aligned}$$

where $c_{\mathbf{h}}$ is a constant term free of the parameters.

Incorporating the priors and marginalizing with respect to $\mathbf{M}$ and $\mathbf{N}$ we

have

$$p(\mathbf{p} \mid \mathbf{v}^*, \mathbf{T}, \mathbf{U}, \gamma^{(1)}, \gamma^{(2)}) \propto \sum_{\mathbf{h} \in \mathcal{H}} \frac{1}{c_{\mathbf{h}}} \prod_{i=1}^{C} \Bigg(p_i^{\sum_{s=1}^{C}\sum_{t=1}^{C} b_{h_{st}i}^{(st)} + \delta - 1} \times$$

$$\frac{\prod_{s=1}^{C} \Gamma(\sum_{t=1}^{C} b_{h_{st}i}^{(st)} + t_{is} + \gamma_i^{(1)}(\epsilon + I(i=s)))}{\Gamma\left(\sum_{s=1}^{C}\left(\sum_{t=1}^{C} b_{h_{st}i}^{(st)} + t_{is}\right) + \gamma_i^{(1)}(C\epsilon + 1)\right)} \times$$

$$\frac{\prod_{t=1}^{C} \Gamma(\sum_{s=1}^{C} b_{h_{st}i}^{(st)} + u_{it} + \gamma_i^{(2)}(\epsilon + I(i=t)))}{\Gamma\left(\sum_{t=1}^{C}\left(\sum_{s=1}^{C} b_{h_{st}i}^{(st)} + u_{it}\right) + \gamma_i^{(2)}(C\epsilon + 1)\right)} \Bigg)$$

$$\propto \sum_{\mathbf{h} \in \mathcal{H}} w_{\mathbf{h}}(\gamma^{(1)}, \gamma^{(2)}, \epsilon) \prod_{i=1}^{C} p_i^{\sum_{s=1}^{C}\sum_{t=1}^{C} b_{h_{st}i}^{(st)} + \delta - 1}$$

where $w_{\mathbf{h}}(\gamma^{(1)}, \gamma^{(2)}, \epsilon)$ is the weight comprising of all the terms not involving p_i's. Now, let $\mathcal{H}^*$ denote the subset of $\mathcal{H}$ such that for all $\mathbf{h}^* = (h_{11}^*, h_{12}^*, \ldots, h_{CC}^*)' \in \mathcal{H}^*$, each index h_{st}^* corresponds to a partition of v_{st} which allocates v_{st} to the s^{th} partition and zero to all the other partitions. Clearly, for any $\mathbf{h}^*$, $\sum_{s=1}^{C}\sum_{t=1}^{C} b_{h_{st}^* i}^{(st)} = \sum_{s=1}^{C}\sum_{t=1}^{C} v_{st} I(i=s) = \sum_{s=1}^{C} I(i=s) v_s = v_i$.

Let ζ denote a generic positive constant which does not depend on ϵ. We absorb terms of the form $\lim_{\epsilon \to 0} \Gamma(x + \mathcal{O}(\epsilon))$ where x is always greater than 1 into ζ, as these limits will be non-zero. Noting that $t_{is} = 0$ if $s \neq i$, for any $\mathbf{h}^* \in \mathcal{H}$ and $\mathbf{h} \in \mathcal{H} \setminus \mathcal{H}^*$, we have

$$\lim_{\epsilon \to 0} \frac{w_{\mathbf{h}}}{w_{\mathbf{h}}^*} = \zeta \prod_{i=1}^{C} \frac{\prod_{s \neq i} \Gamma(\sum_{t=1}^{C} b_{h_{st}i}^{(st)} + \gamma_i^{(1)}\epsilon)}{\prod_{s \neq i} \Gamma(\gamma_i^{(1)}\epsilon)} \frac{\prod_{t \neq i} \Gamma(\sum_{s=1}^{C} b_{h_{st}i}^{(st)} + u_{it} + \gamma_i^{(2)}\epsilon)}{\prod_{t \neq i} \Gamma(\sum_{s=1}^{C} b_{h_{st}^* i}^{(st)} + u_{it} + \gamma_i^{(2)}\epsilon)}$$

Since u_{it}'s are greater than zero and there exists at least one pair (i, s) such that $\sum_{t=1}^{C} b_{h_{st}i}^{(st)} > 0$, the result follows. □

2.9.5 Detailed analysis of the simulation results

In this Section we present a much more thorough analysis of the simulation study, as well as investigate as well as investigate additional methods to generate population-level class probabilities in the target domain.

2.9.5.1 Impact of difference in marginal class distributions between source and target domains

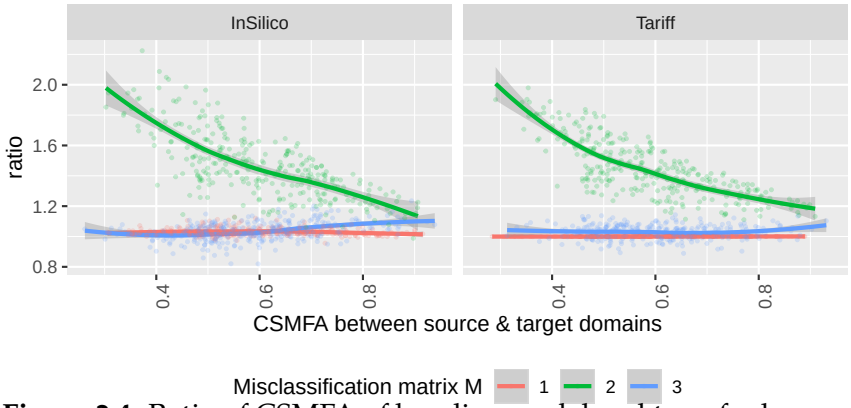


Figure 2.4: Ratio of CSMFA of baseline model and transfer learner

We first investigate how the performance of our Bayesian transfer learning model is impacted by the disparity in class distribution between $\mathcal{D}_S$ and $\mathcal{D}_T$. Figure 2.4 plots the smoothed ratio of the CSMFA of the baseline estimates and their calibrated analogs from our model, as a function of the true CSMFA between the class probabilities $\mathbf{p}_{\mathcal{G}}$ and $\mathbf{p}_{\mathcal{U}}$ in the source and target domains. The left panels correspond to data generated using InSilicoVA and hence assesses the performance of InSilicoVA$_{\mathcal{G}}$ and InSilicoVA$_{BTL}$ by plotting the ratio CSMFA(InSilicoVA$_{BTL}$)/ CSMFA(InSilicoVA$_{\mathcal{G}}$). Similarly, the right panels correspond to data generated using Tariff and compares the estimates from

Tariff$_{\mathcal{G}}$ and Tariff$_{BTL}$. We only present the results for $n = 400$, as we will discuss the role of n in Section 2.9.5.3.

We first note that when data was generated using the misclassification matrix $\mathbf{M}_1 = \mathbf{I}$, the ratio is exactly one. This corroborates the result in Theorem 1 that if A classifies flawlessly in $\mathcal{D}_T$, then the baseline and transfer learning estimates are same. For $\mathbf{M}_3$, i.e. when the misclassification rate is small, the ratio is also close to one with the transfer learning estimate being slightly more accurate in general. For $\mathbf{M}_2$, which portrays the scenario where the the baseline learner trained on source domain is systematically and substantially biased, one can clearly see the benefit of transfer learning. The CSMFA is significantly better after transfer learning. It also nicely shows the utility of transfer learning as a function of $x = CSMFA(\mathbf{p}_{\mathcal{G}}, \mathbf{p}_{\mathcal{U}})$ (on the x-axis). Unsurprisingly, the ratio is decreasing with increasing x. When x is small, i.e., there exists much disparity in the marginal class distributions between the source and target domains, the ratio is close to two, implying that transfer learning yields near 100% gain in accuracy. When x is close to one, the improvement is much less stark, which is expected as in this scenario the class probabilities in the non-local and local populations are almost identical.

2.9.5.2 Biases in estimates of probabilities for each class

We also look at the biases in the estimates of each of the four class probabilities in Figure 2.5. The top and bottom rows correspond to data generated using InSilicoVA and Tariff respectively. The three columns correspond to three

choices of $\mathbf{M}$. We see that there is almost no bias for $\mathbf{M}_1$ for all the methods, for $\mathbf{M}_3$ the baseline Tariff$_{\mathcal{G}}$ estimates are generally unbiased, whereas the baseline InSilicoVA$_{\mathcal{G}}$ show small biases. However, for $\mathbf{M}_2$ we see the substantial biases in the estimates from both the baseline approaches. As expected due to the specification of $\mathbf{M}_2$, the baseline learners underestimate $P(Diarrhea/Dysentry)$ (Cause 2) and $P(Sepsis)$ (Cause 3) while overestimating $P(Pneumonia)$ (Cause 1) and $P(Other)$ (Cause 4) are overestimated. The transfer learning estimates Tariff$_{BTL}$ and InSilicoVA$_{BTL}$ are unbiased for all the settings.

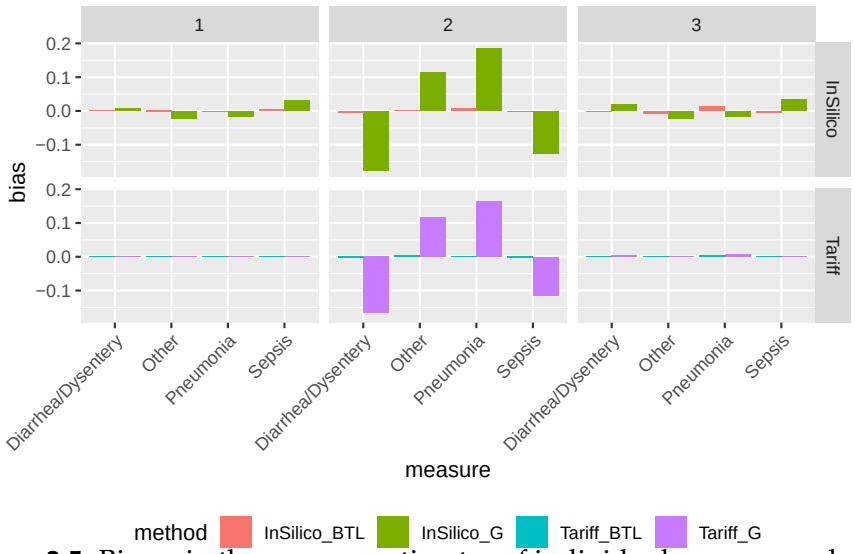


Figure 2.5: Biases in the average estimates of individual cause prevalences

2.9.5.3 Role of limited labeled data in target domain

We now investigate the role of the sample size n and the marginal class distribution $\mathbf{p}_{\mathcal{L}}$ of $\mathcal{L}$. Additionally, as an alternate to our transfer learning approach, we also consider including the local labeled data $\mathcal{L}$ as part of the training data for the CCVA algorithms. So, we have four more methods Tariff$_{\mathcal{L}}$,

Tariff$_{\mathcal{G}\cup\mathcal{L}}$, InSilicoVA$_{\mathcal{L}}$ and InSilicoVA$_{\mathcal{G}\cup\mathcal{L}}$, where the sub-scripts indicate the training data used.

When data is generated using InSilicoVA, Figure 2.6 provides the boxplots of CSMF accuracy of the methods for all the scenarios as a function of n (rows), choice of $\mathbf{M}$ (columns) and ρ — the CSMFA-range between $\mathbf{p}_{\mathcal{L}}$ and $\mathbf{p}_{\mathcal{U}}$ (x-axis in each sub-figure).

We unpack many different conclusions from this Figure. First we look at the performances of InSilicoVA$_{\mathcal{G}}$ and InSilicoVA$_{BTL}$. These two methods was already compared in Section 2.9.5.1, but only for fixed $n = 400$ and averaged across all ρ. Here, further analyzing the performances as a function of n and ρ, we see that the CSMFA of calibrated VA using our model increases with increase in n. Also, there is a drastic gain in precision of the calibrated estimates with the confidence bands shortening with increase in n from 50 to 400. Additionally, we see that the CSMFA for InSilicoVA$_{BTL}$ increases as ρ goes from *low* to *medium* to *high*, although the gain is not as drastic. This indicates that the transfer learning procedure, while being reasonably robust to the value of ρ, does benefit to a small extent from improved concordance between the class probabilities in $\mathcal{L}$ and $\mathcal{U}$. Of course, InSilicoVA$_{\mathcal{G}}$ is not affected by either n or ρ. In general, for $\mathbf{M}_3$, we see that only when both n is small and ρ is *low*, the InSilicoVA$_{\mathcal{G}}$ produces slightly better estimates than InSilicoVA$_{BTL}$. For all other cases, InSilicoVA$_{BTL}$ yields higher or similar CSMF. For $\mathbf{M}_2$, we see InSilicoVA$_{BTL}$ dominates InSilicoVA$_{\mathcal{G}}$ across all scenarios. The gains from increase in n and ρ are evident here as well. Finally, for $\mathbf{M}_1$, the performance of InSilicoVA$_{BTL}$ is identical to InSilicoVA$_{\mathcal{G}}$, as is guaranteed by Theorem 1,

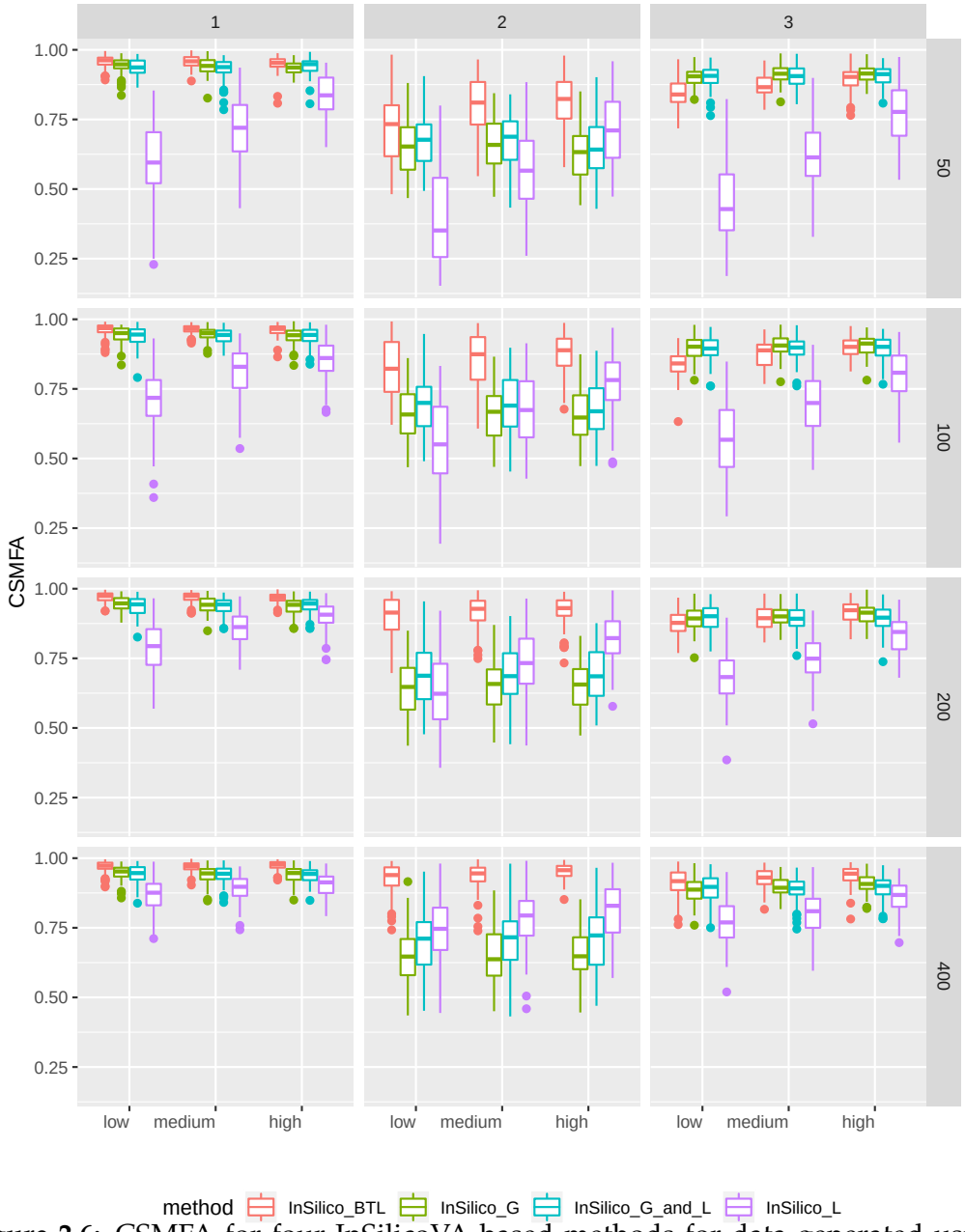


Figure 2.6: CSMFA for four InSilicoVA based methods for data generated using InSilicoVA

and is not affected by n or ρ.

Next, we look at the performance of InSilicoVA$_{\mathcal{L}}$ and InSilicoVA$_{\mathcal{G}\cup\mathcal{L}}$. For $\mathbf{M}_1$ and $\mathbf{M}_3$, InSilicoVA$_{\mathcal{L}}$ performs quite poorly, generally producing the lowest CSMF. InSilicoVA$_{\mathcal{L}}$ is also highly sensitive to both ρ and n, yielding highly variable and inaccurate estimates for *low* ρ and n, and improving sharply as either increases. Only for $\mathbf{M}_2$, for large n or large ρ, it does better than InSilicoVA$_{\mathcal{G}}$. As this setting portrays substantial difference in the conditional distributions between the source and target population, InSilicoVA$_{\mathcal{L}}$, trained on local data, does better. CSMFA from InSilicoVA$_{\mathcal{G}\cup\mathcal{L}}$, which uses both the source and target labeled data in the training, generally lies between the CSMFA from InSilicoVA$_{\mathcal{G}}$ and InSilicoVA$_{\mathcal{L}}$, and is much closer to the former as $\mathcal{G}$ far outnumbers $\mathcal{L}$. Finally, comparing InSilicoVA$_{\mathcal{L}}$ and InSilicoVA$_{\mathcal{G}\cup\mathcal{L}}$ to InSilicoVA$_{BTL}$, we see that the InSilicoVA$_{BTL}$ does substantially better than InSilicoVA$_{\mathcal{L}}$ uniformly across the scenarios, and than InSilicoVA$_{\mathcal{G}\cup\mathcal{L}}$ across all scenarios except when both n is small and ρ is *low*. This shows that with a small labeled dataset in $\mathcal{D}_T$, our transfer learning approach is a more resourceful way of exploiting this limited data and results in more accurate and robust estimates. The analogous results for data generated using Tariff, provided in Figure 2.13 of the supplement, reveals similar trends.

2.9.5.4 Comparison with the naive transfer learning

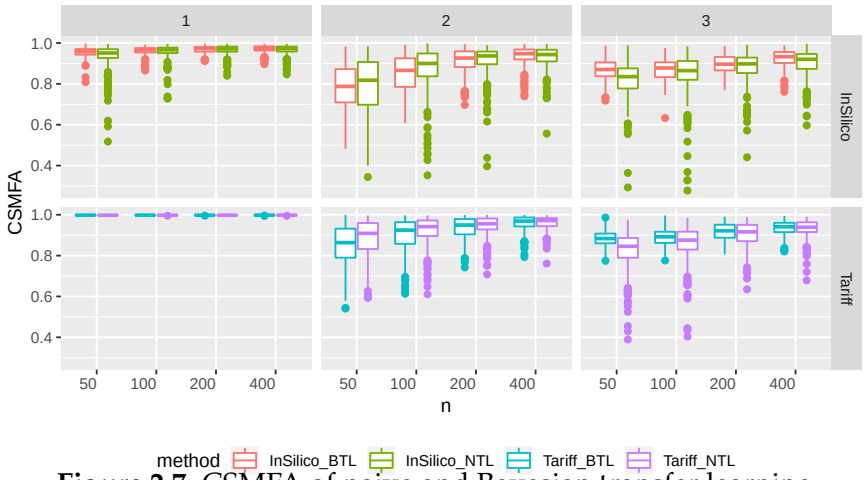


Figure 2.7: CSMFA of naive and Bayesian transfer learning

To understand the importance of the Bayesian shrinkage or regularization used in the the transfer learning, we also compare with the naive transfer learning based on MLE, outlined in Section 2.2. We refer to the naive transfer learning using Tariff and InSilicoVA respectively as Tariff_{NTL} and InSilicoVA_{NTL}. Figure 2.7 compares the CMSFA for the naive and Bayesian regularized transfer learning approaches. Once again, the top and bottom row corresponds to data generated using Tariff and InSilicoVA respectively, the three columns are for three choices of $\mathbf{M}$ and within each setting, we plot the boxplots of CSMFA as a function of n.

We see that, generally the median estimates from the naive approach is similar to the ones produced using the Bayesian regularized analog. However, there is notable difference in the variability of CSMFA, with the naive approach producing a wide range of values with several extreme estimates. The problem is exacerbated for smaller values of n. The results from the Bayesian model

are more stable with uniformly lesser variation across all the settings. It is evident, that in real data analysis, where the truth is unknown, the Bayesian model will be much more reliable than the MLE based solution which seems to be quite likely to yield absurd estimates.

2.9.5.5 Performance of ensemble models

We now analyze the performance of the joint (Ensemble$_J$) and independent (Ensemble$_I$) ensemble transfer learning models introduced in Section 2.3. These models use output from both Tariff and InSilicoVA whereas the single-classifier models Tariff$_{BTL}$ and InSilicoVA$_{BTL}$ only use the output from one CCVA algorithm. For a given dataset, we define

$$\delta = \max(\text{CSMFA}(\text{InSilicoVA}_{BTL}), \text{CSMFA}(\text{Tariff}_{BTL})) -$$

$$\min(\text{CSMFA}(\text{InSilicoVA}_{BTL}), \text{CSMFA}(\text{Tariff}_{BTL})).$$

In other words, δ denotes the difference in CSMFA of the calibrated VA using the most and least accurate classifiers. A small δ implies transfer learning with either of the baseline classifiers yield similar results, whereas larger values of δ clearly insinuate that transfer learning with one of the baseline classifiers is more accurate than the other one. For an ensemble method that aims to guard against inclusion of an inaccurate method, one would expect that CSMFA for the ensemble method should be closer to that of the best performing method. Equivalently, if

$$\nu = CSMFA(Ensemble) - \min(\text{CSMFA}(\text{InSilicoVA}_{BTL}), \text{CSMFA}(\text{Tariff}_{BTL})),$$

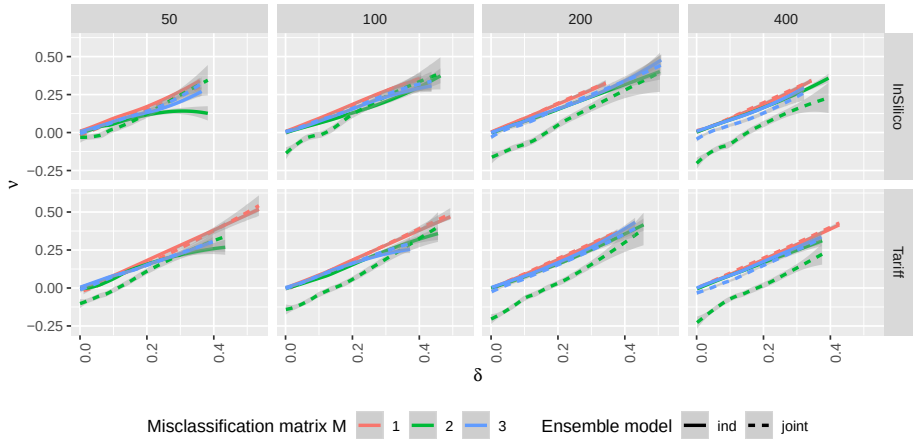


Figure 2.8: Performance of the ensemble models

where Ensemble refers to either Ensemble$_I$ or Ensemble$_J$, then ν should be greater than $\delta/2$.

Figure 2.8 plots ν as a smoothed function of δ. We first note that, for $\mathbf{M}_1$ (red lines), the (ν, δ) curve for the joint sampler nearly coincides with the 45-degree line. Since, in our data generation process, under $\mathbf{M}_1$ one of the classifiers is fully accurate, this empirically verifies the theoretical guarantee in Theorem 2, that in such settings posterior mean of class probabilities from the ensemble approach is same as that from the best classifier. While the independent ensemble model does not enjoy this theoretical property, in practice we see that for $\mathbf{M}_1$, ν is also identical to δ. For $\mathbf{M}_2$ and $\mathbf{M}_3$, across all scenarios, ν is close to δ, i.e. estimates from both the Ensemble$_J$ and Ensemble$_I$ models generally aligns much closer to the best performing single-classifer transfer learner. There are no significant trends with respect to either the size of $\mathcal{L}$ (n) or the data generating algorithm – InSilicoVA (top-row) and Tariff (bottom-row). The Ensemble$_I$ model seems to do slightly better than the joint model.

Since, it is also the faster model, we only use this version of the ensemble model for subsequent analysis. The performance of the ensemble samplers is quite reassuring especially for larger δ, as it demonstrates the robustness to inclusion of a bad method via averaging over multiple methods. As ν seems to be substantially greater than $\delta/2$ for most of the curves, it also shows why our model based method averaging is superior to simply taking average of the estimated class-probabilities from the different methods, which is much more affected by the worst method.

2.9.5.6 Informative shrinkage

If we have prior knowledge on $\mathbf{M}$, we can use this for informative shrinkage, rather than shrinking towards the source predictor. For example, when the true matrix is $\mathbf{M}_2$, if we assume that it was known apriori that label 2 is often misclassified as label 1, and label 3 is often misclassified as label 4, then instead of shrinking $\mathbf{M}$ towards the identity matrix, we can shrink $\mathbf{M}$ towards transition matrices of the form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ m_{21} & m_{22} & 0 & 0 \\ 0 & 0 & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

To do this *informative shrinkage*, we can let

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and use an informed prior on $\mathbf{M}$ such that

$$\mathbf{M}_{i*} \overset{ind}{\sim} \text{Dirichlet}(\gamma_i(\mathbf{X}_{i*} + \epsilon\mathbf{1})), i = 1, 2, 3, 4$$

This prior would reflect our knowledge of which causes are likely to be misclassified by the algorithm. We then modify our Gibbs updates as follows:

$$\mathbf{M}_{i*} \mid \cdot \sim Dirichlet(\mathbf{B}_{i*} + \mathbf{T}_{i*} + \lambda_i(\mathbf{X}_{i*} + \epsilon\mathbf{1}))$$

$$p(\gamma_i \mid \cdot) \propto \frac{\Gamma(C\gamma_i\epsilon + \gamma_i \cdot \sum_j \mathbf{1}(\mathbf{X}_{ij} = 1))}{\prod_j \Gamma(\gamma_i\epsilon + \gamma_i\mathbf{1}(\mathbf{X}_{ij} = 1))} \gamma_i^{\alpha-1} \exp(-\beta\gamma_i) \prod_j m_{ij}^{\gamma_i\epsilon+\gamma_i 1(\mathbf{X}_{ij}=1)}$$

While the choice of prior is less likely to affect the results with a larger calibration set size, we can compare the CSMFA when using a smaller calibration set size in Figure 2.9 below.

We see that with a sample size of 50 for our $\mathcal{L}$, using the informed prior on $\mathbf{M}$ leads to improved CSMFA. When the sample size for $\mathcal{L}$ grows to 100, there is still some improvement in CSMFA with informed shrinkage when the data is generated using InSilicoVA, and the performance is nearly identical between the two models when the data is generated by Tariff.

2.9.5.7 Individual level classification

As mentioned earlier, predicting individual classes is not our primary goal. Nonetheless, we have outlined a simple way to obtain individual predictions using our transfer learning model. Here we compared its accuracy using the

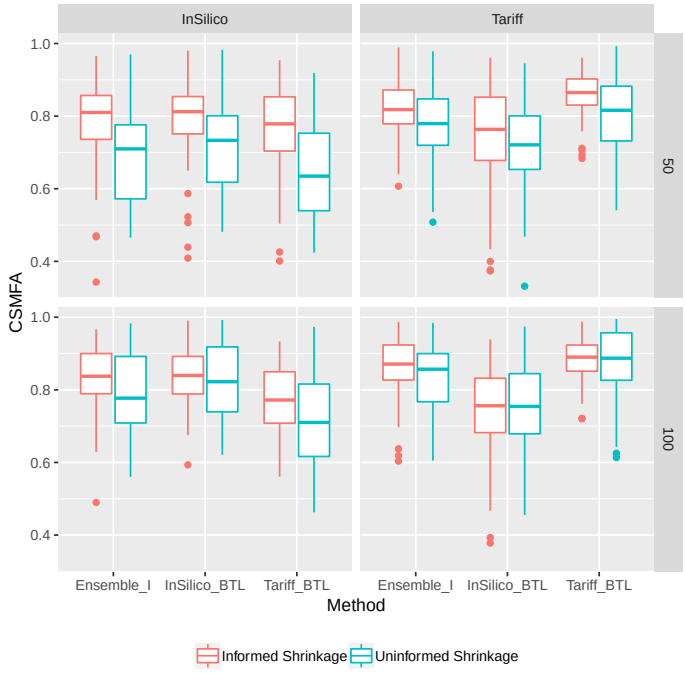


Figure 2.9: Comparison between informative and non-informative (default) shrinkage.

Chance Corrected Concordance (Murray et al., 2011a) defined as

$$\text{CCC} = \frac{1}{C}\sum_{i=1}^{C}\frac{\frac{TP_i}{TP_i+TN_i} - \frac{1}{N}}{1-\frac{1}{N}}$$

where TP_i and TN_i denote the true positive and true negative rates for class i. We only analyze the case when when the data is generated using InSilicoVA (Figure 2.10). The roles are simply reversed when data is generated using Tariff (Figure 2.14). We see in Figure 2.10 that CCC for InSilicoVA$_{\mathcal{G}}$ and InSilicoVA$_{BTL}$ are better than those of Tariff$_{\mathcal{G}}$ and Tariff$_{BTL}$ respectively. This is expected as analyzing data using the true model is expected to perform better than the misspecified model. CCC from the transfer learning (InSilicoVA$_{BTL}$) and baseline (InSilicoVA$_{\mathcal{G}}$) versions of the same CCVA algorithm, which was

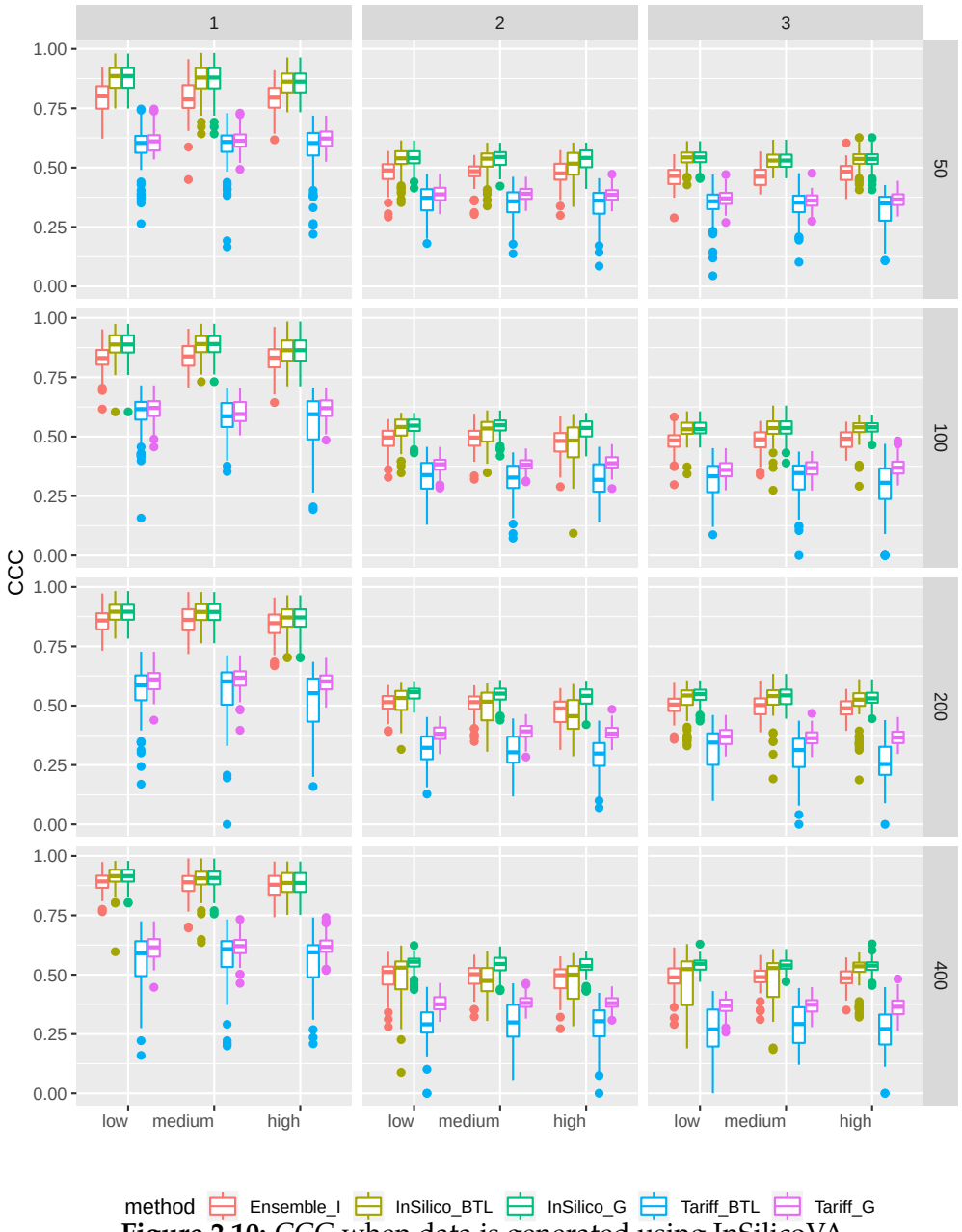


Figure 2.10: CCC when data is generated using InSilicoVA

used in data generation, were similar. For the misspecified model, the baseline Tariff$_{\mathcal{G}}$ produced slightly better CCC than the transfer learning Tariff$_{BTL}$, although this improvement in performance is minor. Overall, these results indicate that if individual prediction is of interest, then perhaps more advanced methods need to be considered than the simple approach we have outlined. However, even using our crude approach, we see that the ensemble model (Ensemble$_I$) produces CCC closer to InSilicoVA$_{BTL}$ and InSilicoVA$_{\mathcal{G}}$, and much better than the CCC obtained by both Tariff$_{BTL}$ and Tariff$_{\mathcal{G}}$. This once again furnishes evidence of the robust performance of the ensemble model, and in practice, when we will not know which algorithm works best, using the ensemble model will safeguard against choosing a bad algorithm.

2.9.6 Comparing marginal symptom distributions between $\mathcal{L}$ and $\mathcal{U}$

To show that our method also does not assume similar symptom marginal distributions between $\mathcal{L}$ and $\mathcal{U}$, in Figure 2.11 we plot the proportion of presence ("Yes") of each symptom in $\mathcal{U}$ and $\mathcal{L}$ (for 10 randomly selected samples of $\mathcal{L}$) in India and Tanzania for this analysis. We see that while many symptoms are rare in both $\mathcal{L}$ and $\mathcal{U}$ (clustering near (0,0)), the marginal distributions of the symptoms do not have to match between $\mathcal{U}$ and $\mathcal{L}$, with the symptom proportion in $\mathcal{L}$ varying considerably on both sides (up to $\pm$15%) than the analogous quantity in $\mathcal{U}$)

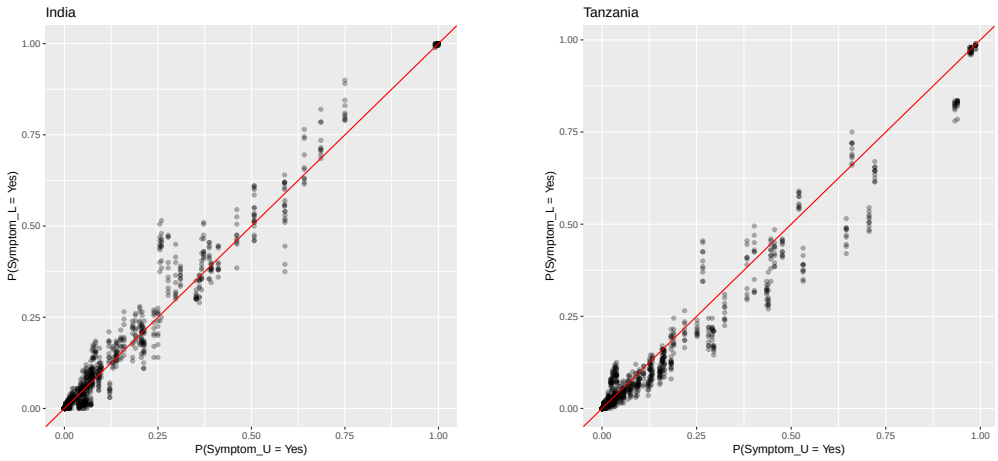


Figure 2.11: Scatterplot of the symptom proportions in $\mathcal{U}$ and 10 randomly sampled choices of $\mathcal{L}$. The red line is the $x = y$ line.

2.9.7 Impact of number of cause categories for PHMRC analysis

To investigate the effect of adding more causes of death on the transfer learning CSMFA, we added "Malaria" and "Sepsis", which were part of the "Other Infectious" category, as individual causes. Due to the nature of the CSMFA metric, it is difficult to directly compare accuracy on estimating a probability vector of length 5 versus a probability vector of length 7. Hence, after getting the transfer learning CSMF estimates for the 7 cause categories, we aggregated the 7 cause CSMFs back to the orignal 5 cause CSMFs, i.e., we added the CSMF estimates for "Malaria" and "Sepsis" to the CSMF estimate for "Other Infectious", so that we could fairly compare the CSMFA when using 5 versus 7 causes.

Looking at Figure 2.12, we see that there is actually very little change in the CSMFA when using individual algorithms. This would indicate that when we only are using one algorithm, the additional causes are not causing

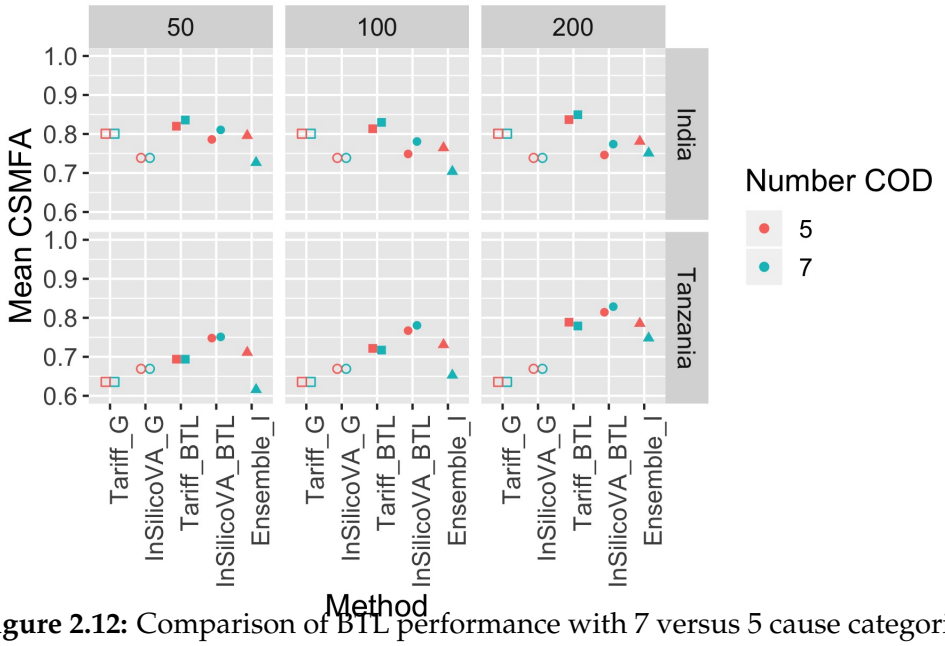


Figure 2.12: Comparison of BTL performance with 7 versus 5 cause categories

substantial shrinkage in the estimates of $\mathbf{M}$. We would expect that as the number of causes grows even larger and the size of $\mathcal{L}$ is small, there are fewer number of samples per cause category leading to more shrinkage towards the source predictor and hence worse performance. We only see this for the ensemble model and for sample sizes 50 and 100 when we add additional causes. This is also most likely due to the fact that the ensemble method requires estimating substantially more parameters with an increased number of causes, as compared to the individual algorithm transfer learning. As the sample size of $\mathcal{L}$ grows larger, we are able to better estimate this increased number of parameters.

2.9.7.1 Additional figures

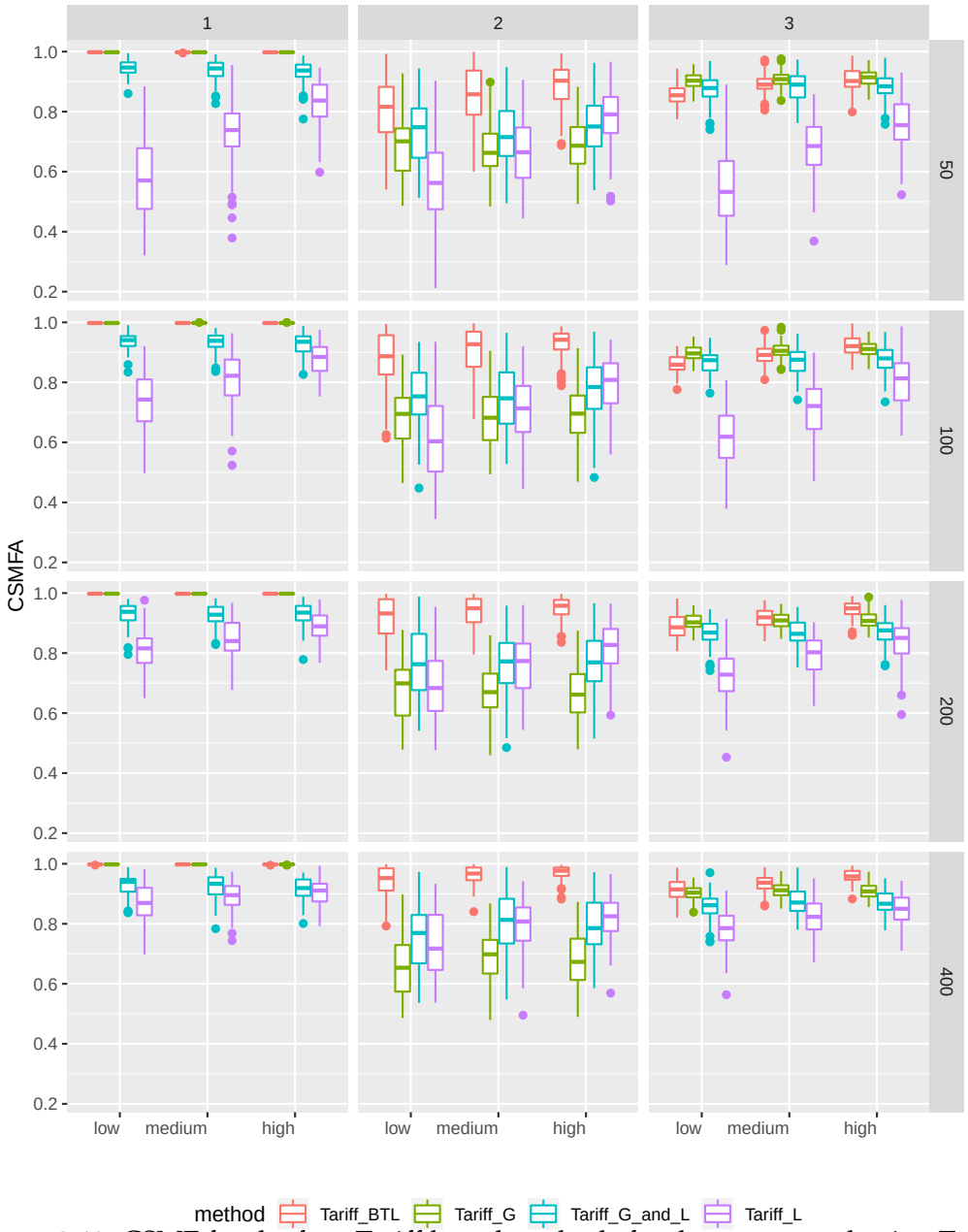


Figure 2.13: CSMF for the four Tariff-based methods for data generated using Tariff

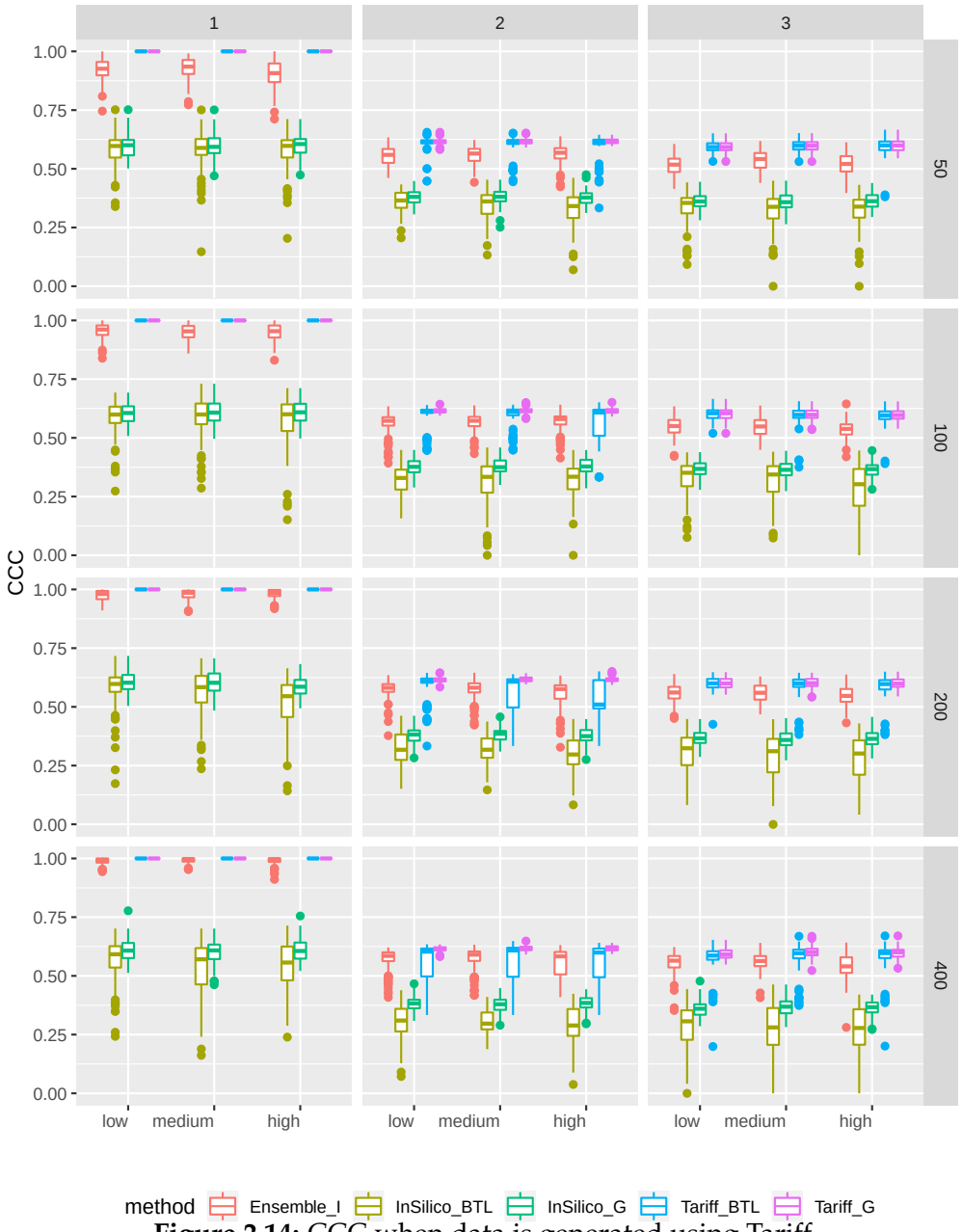


Figure 2.14: CCC when data is generated using Tariff

References

Shimodaira, Hidetoshi (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of statistical planning and inference* 90.2, pp. 227–244.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang (2016). "A survey of transfer learning". In: *Journal of Big Data* 3.1, p. 9.

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". In: *IEEE Trans. on Knowl. and Data Eng.* 22.10, pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. URL: http://dx.doi.org/10.1109/TKDE.2009.191.

Chattopadhyay, Rita, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye (2012). "Multisource domain adaptation and its application to early detection of fatigue". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6.4, p. 18.

Oquab, Maxime, Leon Bottou, Ivan Laptev, and Josef Sivic (2014). "Learning and transferring mid-level image representations using convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724.

Dai, Wenyuan, Qiang Yang, Gui-Rong Xue, and Yong Yu (2007). "Boosting for Transfer Learning". In: *International Conference on Machine Learning, Corvallis, OR*.

Yao, Yi and Gianfranco Doretto (2010). "Boosting for transfer learning with multiple sources". In: *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, pp. 1855–1862.

Daumé III, Hal (2009). "Frustratingly easy domain adaptation". In: *arXiv preprint arXiv:0907.1815*.

Pan, Sinno Jialin, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen (2010). "Cross-domain sentiment classification via spectral feature alignment". In: *Proceedings of the 19th international conference on World wide web*. ACM, pp. 751–760.

AbouZahr, Carla, Don De Savigny, Lene Mikkelsen, Philip W Setel, Rafael Lozano, Erin Nichols, Francis Notzon, and Alan D Lopez (2015). "Civil registration and vital statistics: progress in the data revolution for counting and accountability". In: *The Lancet* 386.10001, pp. 1373–1385.

Allotey, Pascale A, Daniel D Reidpath, Natalie C Evans, Nirmala Devarajan, Kanason Rajagobal, Ruhaida Bachok, Kridaraan Komahan, and SEACO Team (2015). "Let's talk about death: data collection for verbal autopsies in a demographic and health surveillance site in Malaysia". In: *Global health action* 8.1, p. 28219.

Soleman, Nadia, Daniel Chandramohan, and Kenji Shibuya (2006). "Verbal autopsy: current practices and challenges". In: *Bulletin of the World Health Organization* 84, pp. 239–245.

James, Spencer L, Abraham D Flaxman, and Christopher JL Murray (2011). "Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies". In: *Population Health Metrics* 9.1, p. 31.

Serina, Peter, Ian Riley, Andrea Stewart, Spencer L James, Abraham D Flaxman, Rafael Lozano, Bernardo Hernandez, Meghan D Mooney, Richard Luning, Robert Black, et al. (2015). "Improving performance of the Tariff Method for assigning causes of death to verbal autopsies". In: *BMC medicine* 13.1, p. 291.

Byass, Peter, Daniel Chandramohan, Samuel J Clark, Lucia D'ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. (2012). "Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool". In: *Global health action* 5.1, p. 19281.

McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark (2016). "Probabilistic cause-of-death assignment using verbal autopsies". In: *Journal of the American Statistical Association* 111.515, pp. 1036–1049.

King, Gary, Ying Lu, et al. (2008). "Verbal autopsy methods with multiple causes of death". In: *Statistical Science* 23.1, pp. 78–91.

Kalter, Henry D, Abdoulaye-Mamadou Roubanatou, Alain Koffi, and Robert E Black (2015). "Direct estimates of national neonatal and child cause–specific mortality proportions in Niger by expert algorithm and physician–coded analysis of verbal autopsy interviews". In: *Journal of global health* 5.1.

Li, Zehang, Tyler McCormick, and Sam Clark (2018c). *Tariff: Replicate Tariff Method for Verbal Autopsy*. URL: https://CRAN.R-project.org/package=Tariff.

Li, Zehang, Tyler McCormick, and Sam Clark (2018a). *InSilicoVA: Probabilistic Verbal Autopsy Coding with 'InSilicoVA' Algorithm*. URL: https://CRAN.R-project.org/package=InSilicoVA.

Thomas, Jason, Zehang Li, Tyler McCormick, Sam Clark, and Peter Byass (2018). *InterVA5: Replicate and Analyse 'InterVA5'*. URL: https://CRAN.R-project.org/package=InterVA5.

Li, Zehang, Tyler McCormick, and Sam Clark (2018b). *openVA: Automated Method for Verbal Autopsy*. URL: https://CRAN.R-project.org/package=openVA.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Minsky, Marvin (1961). "Steps toward artificial intelligence". In: *Proceedings of the IRE* 49.1, pp. 8–30.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Flaxman, Abraham D, Alireza Vahdatpour, Sean Green, Spencer L James, and Christopher JL Murray (2011). "Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards". In: *Population health metrics* 9.1, p. 29.

Miasnikof, Pierre, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha (2015). "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths". In: *BMC medicine* 13.1, p. 286.

Koopman, Bevan, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, and Sarah Thackway (2015). "Automatic classification of diseases from free-text death certificates for real-time surveillance". In: *BMC medical informatics and decision making* 15.1, p. 53.

Byass, Peter (2016). "Minimally invasive autopsy: A new paradigm for understanding global health?" In: *PLoS medicine* 13.11, e1002173.

Murray, CJL, AD Lopez, R Black, and et al. (2011b). "Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets". In: *Population Health Metrics*, pp. 9–27.

Flaxman, Abraham D, Jonathan C Joseph, Christopher JL Murray, Ian Douglas Riley, and Alan D Lopez (2018). "Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards". In: *BMC medicine* 16.1, p. 56.

Leitao, Jordana, Nikita Desai, Lukasz Aleksandrowicz, Peter Byass, Pierre Miasnikof, Stephen Tollman, Dewan Alam, Ying Lu, Suresh Kumar Rathi, Abhishek Singh, et al. (2014). "Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low-and middle-income countries: systematic review". In: *BMC medicine* 12.1, p. 22.

Long, Mingsheng, Jianmin Wang, Guiguang Ding, Sinno Jialin Pan, and S Yu Philip (2014). "Adaptation regularization: A general framework for transfer learning". In: *IEEE Transactions on Knowledge and Data Engineering* 26.5, pp. 1076–1089.

Polson, Nicholas G, James G Scott, and Jesse Windle (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American statistical Association* 108.504, pp. 1339–1349.

Murray, Christopher JL, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez (2011a). "Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies". In: *Population health metrics* 9.1, p. 28.

Chapter 3

A Transformation-free Linear Regression for Compositional Outcomes and Predictors

3.1 Introduction

Compositional data, also referred to as fractional data (Mullahy, 2015; Murteira and Ramalho, 2016), consist of vectors constrained to lie in the unit simplex, $\mathbb{S}^D$, where $\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D)' | x_j \geq 0, i = j, \ldots, D; \sum_{i=j}^{D} x_j = 1\}$. Compositional data appear in many fields, such as econometrics (Papke and Wooldridge, 1996), geochemistry (Templ, Filzmoser, and Reimann, 2008), physical activity research (Dumuid et al., 2018), microbiome analysis (Lin et al., 2014), and nutritional epidemiology (Leite, 2016).

Depending on the application, compositional data may appear as an explanatory variable (Hron, Filzmoser, and Thompson, 2012; McGregor et al., 2019; Dumuid et al., 2018), as an outcome of interest (Papke and Wooldridge, 1996; Mullahy, 2015; Egozcue et al., 2012; Hijazi and Jernigan, 2009), or both (Wang et al., 2013; Chen, Zhang, and Li, 2017; Alenazi, 2019). While there

has been much attention placed on the first two cases, little work has been done on creating simple and interpretable models for the last case. Examples of problems with both compositional outcomes and explanatory variables include relating the percentage of males and females with different education levels across countries (Filzmoser, Hron, and Templ, 2018), modeling the relationship between age structure and consumption structure across economic areas (Chen, Zhang, and Li, 2017), and understanding how different methods for estimating the composition of white blood cell types are related (Aitchison, 1986; Alenazi, 2019).

All current methods developed specifically for problems where both the outcome and the explanatory variable are compostional require transforming the compositional data. Chen, Zhang, and Li (2017) transforms both the response and explanatory compositional variables, while Alenazi (2019) transforms just the compositional explanatory variable. Transformation based models limit interpretability (Morais, Thomas-Agnan, and Simioni, 2018), especially when complex, but commonly used transformations such as the isometric log-ratio (ILR) transformation (Egozcue et al., 2003) are used. Furthermore, many transformations do not allow for compositional data with 0s and 1s (Filzmoser, Hron, and Templ, 2018).

In this manuscript, we postulate a simple estimating equation that directly relates the expected value of the compositional outcome as a linear function of the compositional explanatory variable. Our approach does not require any transformation of the data and naturally accommodates 0s and 1s, thus treating data on the interior of the simplex the same as data on the boundary.

By linearly relating the outcome and explanatory variables, the parameters in our model are easily interpretable, unlike transformation based compositional regression models. We develop an expectation-maximization (EM) (Dempster, Laird, and Rubin, 1977) algorithm for fast and accurate parameter estimation via constrained maximization of the quasi-likelihood that respects the unit sum nature of the compositional data. We present simulation results comparing the models for compositional data under a variety of data generating mechanisms. We also present a permutation-based test for assessing whether or not there exists a linear dependency between the outcome and explanatory variables, and evaluate the operating characteristics of this test via simulation. Finally, we demonstrate the utility of our model with two data analyses from education and medical research.

3.2 Review of Transformation Based Compositional Regression Models

Current models for problems with compositional outcomes and explanatory variables rely on transforming the compositional data from $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$. The recommended transformation for compositional data is the ILR transformation (Egozcue et al., 2003; Hron, Filzmoser, and Thompson, 2012; Filzmoser, Hron, and Templ, 2018), where for $\mathbf{z} \in \mathbb{S}^D$

$$ilr(\mathbf{z})_j = \sqrt{\frac{D-j}{D-j+1}} ln \left(\frac{z_j}{\left(\prod_{k=j+1}^{D} z_k \right)^{\frac{1}{D-j}}} \right), \; j = 1, \ldots, D-1.$$

The mathematical advantage of using the ILR transformation over more simple transformations, such as the additive log-ratio (ALR) or centered log-ratio (CLR) (Aitchison, 1986), is that the vector $ilr(\mathbf{z})$ can be used as covariates in a standard linear regression model without having to constrain the regression coefficients (Hron, Filzmoser, and Thompson, 2012).

The model presented by Chen, Zhang, and Li (2017) assumes that for an outcome $\mathbf{y} \in \mathbb{S}^{D_r}$ and explanatory variable $\mathbf{x} \in \mathbb{S}^{D_s}$, where D_r is not necessarily equal to D_s, that

$$E[ilr(\mathbf{y})_k|\mathbf{x}] = \beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} ilr(\mathbf{x})_j,\ k = 1, \ldots, D_r - 1. \tag{3.1}$$

Hence, β_{11} has an interpretation as the effect of increasing the relative value of x_1 by 1 compared to the rest of $\mathbf{x}$, holding the ratios between the other components of $\mathbf{x}$ constant, on the change of the relative value of y_1 compared to the rest of $\mathbf{y}$; the other regression coefficients have no meaningful interpretation (Hron, Filzmoser, and Thompson, 2012; Chen, Zhang, and Li, 2017). To obtain the effects of relative changes of each part of $\mathbf{x}$ on $\mathbf{y}$, one must use the permutation operation,

$$\mathbf{z}^l = (z_l, z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_D),$$

and estimate $D_r \cdot D_s$ separate models where

$$E[ilr(\mathbf{y}^{l_1})_k] = \beta_{0k}^{(l_1,l_2)} + \textstyle\sum_{j=1}^{D_s-1} \beta_{jk}^{(l_1,l_2)} ilr(\mathbf{x}^{l_2})_j,\ k = 1, \ldots, D_r - 1,\ l_1 = 1, \ldots, D_r,\ l_2 = 1, \ldots, D_s. \tag{3.2}$$

The coefficients of interest would then be $\beta_{11}^{(l_1,l_2)}$ for each combination of

l_1 and l_2 (Chen, Zhang, and Li, 2017; Filzmoser, Hron, and Templ, 2018). As parameter estimation is performed using standard maximum likelihood for linear regression models, this procedure is not computationally expensive. However, using multiple versions of a model to obtain a set of coefficients that cannot be interpreted jointly is undesirable. There are two additional downsides. First, the ILR transformation does not allow for 0s in the compositional data. If either $\mathbf{x}$ or $\mathbf{y}$ are categorical, the ILR transformation framework can not be used, even though categorical variables are still in the unit simplex. Second, the coefficients of interest can only be vaguely interpreted in terms of changes in the relative values of each part of the compositional data to the geometric mean. This model does not permit for simple interpretation of the coefficients in terms of the direct effect of changing the value of $\mathbf{x}$ within the simplex on the expected value of $\mathbf{y}$ in the simplex (Morais, Thomas-Agnan, and Simioni, 2018). The lack of a simple interpretation for the coefficients in (3.2) have forced practitioners to instead rely on graphical techniques to display the estimated response surface of $\mathbf{y}$ as a function of $\mathbf{x}$ (Nguyen et al., 2018).

Alenazi (2019) takes a different approach to compositional regression, as only the explanatory compositional variable $\mathbf{x}$ is transformed. While Alenazi (2019) is more interested in prediction accuracy than interpretation and uses a complex principal components based transformation, one can use any transformation t (e.g., the ILR transformation). The assumed regression model is the multinomial logit specification (Papke and Wooldridge, 1996; Mullahy, 2015; Murteira and Ramalho, 2016):

$$
\begin{aligned}
E[y_k|\mathbf{x}] &= \frac{\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)}{1 + \sum_{k=1}^{D-1} \left[\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)\right]}, \; k = 1, \ldots, D_r - 1 \\
E[y_{D_r}|\mathbf{x}] &= \frac{1}{1 + \sum_{k=1}^{D-1} \left[\exp(\beta_{0k} + \sum_{j=1}^{D_s-1} \beta_{jk} t(\mathbf{x})_j)\right]}.
\end{aligned}
\tag{3.3}
$$

Murteira and Ramalho (2016) discuss both quasi-maximum and maximum likelihood (QML and ML) methods for estimation of the coefficients. However, Alenazi (2019) uses a QML method which allows for 0 values in $\mathbf{y}$ (Papke and Wooldridge, 1996; Mullahy, 2015; Murteira and Ramalho, 2016), and does not make any distributional assumptions about $\mathbf{y}$.

Despite this method allowing for potential 0s in $\mathbf{y}$ (and in $\mathbf{x}$ if one uses a transformation that allows for 0s, such as the α-transformation (Tsagris, 2015)), the regression coefficients are still only interpretable in terms of effects of changing a transformed version of $\mathbf{x}$ on $\log\left(\frac{E[y_j]}{E[y_{D_r}]}\right)$. In order to interpret the model in terms of changes within the simplex, one would again need to resort to graphical techniques.

3.3 Direct Regression of Compositional Variables on the Simplex

Section 3.2 showed that current models for regressing a compositional outcome on a compositional explanatory variable are difficult to interpret due to modeling transformed versions of the compositional data. To create an interpretable model for this class of problems, we want to directly model the

expected value of $\mathbf{y}$ as a linear function of $\mathbf{x}$. This is achieved through the following linear model:

$$E[\mathbf{y}|\mathbf{x}] = \sum_{j=1}^{D_s} x_j \mathbf{b}_j, \tag{3.4}$$

where $\mathbf{b}_j$'s are D_y-dimensional vectors. Letting $\mathbf{B}$ represent the matrix with the jth row $\mathbf{B}_{j*} = \mathbf{b}_j^{'}$, we can rewrite the model in (3.4) as

$$E[\mathbf{y}|\mathbf{x}] = \mathbf{B}^{'}\mathbf{x}\,. \tag{3.5}$$

Because $\mathbf{y}$ compositional, we require that $\sum_{k=1}^{D_r} E[y_k|\mathbf{x}] = 1$. To adhere to the unit sum restriction, we take advantage of the fact that $\mathbf{x}$ is also compositional. Hence, it suffices to constrain $\mathbf{B}$ to be a Markov (transition) matrix with non-negative entries and rows summing to 1, i.e.,

$$\mathbf{B} \in \{\mathbb{R}^{D_s \times D_r} | B_{jk} \geq 0, \sum_{k=1}^{D_r} \mathrm{B}_{jk} = 1 \text{ for } j = 1, \ldots, D_s\}\,.$$

This transformation-free model allows 0s and 1s in both x and y as (3.5) is well-defined for entire x- and y-simplexes including the boundaries. The model allows for direct interpretation of the association between $\mathbf{x}$ and $E[\mathbf{y}]$ in terms of the regression coefficient matrix $\mathbf{B}$. If x_j increases by $\Delta \in (0, 1 - x_j]$, at the expense of x_k decreasing by Δ (assuming $x_k \geq \Delta$) and holding the rest of $\mathbf{x}$ constant, the expected change in $E[\mathbf{y}]$ is expressed as $\Delta(\mathbf{B}_{j*} - \mathbf{B}_{k*})$. This interpretation respects the fact that increasing one part of $\mathbf{x}$ necessarily involves the trade-off of decreasing at least one other part of $\mathbf{x}$. For example, if $\mathbf{x}$ represents the proportion of each day spent on different activities such

as sleep, physical activity, and sedentary time, we may be interested in how components of a compositional $\mathbf{y}$ are expected to change when we increase physical activity and decrease sedentary time. We also may be interested in how this compares to the change of $\mathbf{y}$ when we instead increase physical activity at the expense of sleep (Dumuid et al., 2018). Another example application where this interpretation is useful is in marketing, where teams may want to know whether to increase the percentage of expenditure on television advertisements at the expense of radio advertisements or press advertisements in order to best increase their market share (Morais, Thomas-Agnan, and Simioni, 2018). Furthermore, our model allows us to directly see how $\mathbf{y}$, rather than some transformed version of $\mathbf{y}$, is associated with $\mathbf{x}$.

In addition to the simple interpretation, the direct regression model in (3.4) exhibits other convenient statistical properties. First, consider the case when two rows, j_1 and j_2, of $\mathbf{B}$ are equal. This implies that increasing x_{j_1} at the expense of x_{j_2} does not change $E[\mathbf{y}]$. We then have

$$\begin{aligned} E[\mathbf{y}|\mathbf{x}] &= \sum_{j \neq j_1, j_2}^{D_s} x_j \mathbf{b}_j + x_{j_1} \mathbf{b}_{j_1} + x_{j_2} \mathbf{b}_{j_2} \\ &= \sum_{j \neq j_1, j_2}^{D_s} x_j \mathbf{b}_j + \mathbf{b}_{j_1} (x_{j_1} + x_{j_2}) \,, \end{aligned} \tag{3.6}$$

which shows that we can treat the combined categories $x_{j_1} + x_{j_2}$ as a single category. This not only simplifies interpretation of the direct regression model, but also means that there is one less row of $\mathbf{B}$ to estimate.

Similarly, the direct regression model can easily accommodate combining

categories y_{k_1} and y_{k_2}. The direct regression model implies that

$$E[y_{k_1} + y_{k_2}|\mathbf{x}] = \sum_{j=1}^{D_s} B_{jk_1} x_j + \sum_{j=1}^{D_s} B_{jk_2} x_j$$

$$= \sum_{j=1}^{D_s} (B_{jk_1} + B_{jk_2}) x_j.$$

Thus, conditional expectations of linear combinations of $\mathbf{y}$ can be obtained through adding columns of $\mathbf{B}$. Rather than having to perform separate regressions using different linear combinations of the outcome, practitioners can simply perform one regression using the full outcome, and obtain linear combinations of $\mathbf{B}$ post-hoc.

Because $\mathbf{B}$ is a Markov matrix, the rows of $\mathbf{B}$ are themselves members of $\mathbb{S}^{D_r}$. If we let $x_j = 1$, which means that $\mathbf{x}$ is in the jth corner of $\mathbb{S}^{D_s}$, (3.4 shows that $E[\mathbf{y}|x_j = 1] = \mathbf{b}_j$. Thus, $\mathbf{B}_{j*}$ is equivalent to $E[\mathbf{y}]$ when $x_j = 1$. For the case when $D_r = 3$, this means we can actually visualize the coefficients themselves using a ternary diagram (Hamilton and Ferry, 2018). Consider the following two values of $\mathbf{B}$:

$$\mathbf{B}^{(1)} = \begin{pmatrix} .90 & .05 & .05 \\ .05 & .90 & .05 \\ .05 & .05 & .90 \end{pmatrix}; \ \mathbf{B}^{(2)} = \begin{pmatrix} .40 & .30 & .30 \\ .30 & .40 & .30 \\ .30 & .30 & .40 \end{pmatrix}$$

$\mathbf{B}^{(1)}$ represents the setting when $\mathbf{y}$ and $\mathbf{x}$ are highly correlated, while $\mathbf{B}^{(2)}$ represents the setting when $\mathbf{y}$ and $\mathbf{x}$ are weakly correlated. This interpretation is derived directly from the simple analytic interpretation of the direct regression model in (3.4). This interpretation is also seen through plotting the rows of these two matrices in a ternary diagram, as in Figure 3.1. Each

number in the plot corresponds to a row in the two values of $\mathbf{B}$. The plot of $\mathbf{B}^{(1)}$ shows that $E[\mathbf{y}]$ substantially changes with $\mathbf{x}$, as changes in $E[\mathbf{y}]$ with $\mathbf{x}$ can be expressed as scaled differences in the rows of $\mathbf{B}$. However, the plot of $\mathbf{B}^{(2)}$ shows much smaller changes for $E[\mathbf{y}]$ with $\mathbf{x}$. Confidence regions for each row of $\mathbf{B}$ can also be plotted within the diagram. We demonstrate this in the example in Section 3.7.1.

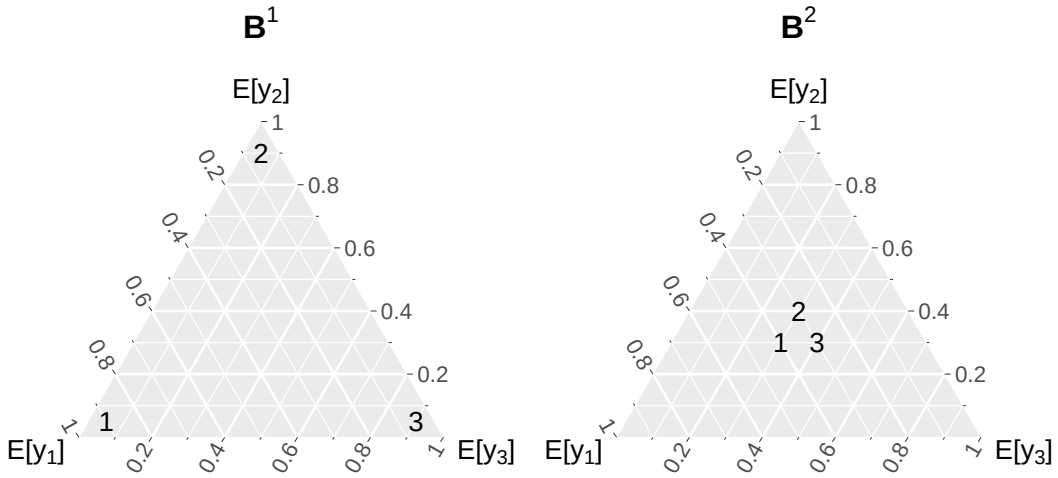


Figure 3.1: Visualization of the coefficients $\mathbf{B}$. For a number j, the point plots $\mathbf{B}_{j*}$ within a ternary diagram.

We note that the models of Chen, Zhang, and Li (2017) and Alenazi (2019) models have some advantages over our simple and direct model, most notably the ability to include multiple confounding covariates of mixed variable type in the model, and we present a full comparison of the properties of each model in Table 1. However, the simple interpretation of the direct regression model stands in stark contrast to the vague interpretation of the coefficients in the ILR model or any model which transforms $\mathbf{y}$ and/or $\mathbf{x}$. The interpretation of

Properties	Direct Regression	ILR transformation (Chen, Zhang, and Li, 2017)	Multinomial logit (Alenazi, 2019)
Transformation-free	✓	✗	✗
Accommodates 0s and 1s in both outcome and predictor compositions	✓	✗	✓
Coefficients interpreted in terms of changes of $\mathbf{y}$ in the simplex	✓	✗	✗
Only requires running 1 model, instead of $D_r \cdot D_s$ models	✓	✗	✗
Coefficients interpreted in terms of changes of log ratios of $\mathbf{y}$	✗	✓	✓
Can be extended to include multiple covariates that may be compositional, continuous, or discrete	✗	✓	✓

Table 3.1: Comparison of properties between the three compositional regression models. A ✓ indicates that a model has the given property, while a ✗ indicates that a model does not have the given property.

B is simple to communicate to non-statisticians without graphical techniques, does not require familiarity with the compositional transformations, and only requires estimating one single model for $E[\mathbf{y}|\mathbf{x}]$, rather than $D_r \cdot D_s$ models. The direct regression model also seamlessly permits 0s and 1s in both $\mathbf{x}$ and $\mathbf{y}$, leading to the sub-cases of interest presented in Sections 3.3.1 and 3.3.2.

3.3.1 Categorical covariates

For each observation i, assume that the covariate of interest is in whether or not the observation belongs to one of $j = 1, \ldots, D_s$ groups. If observation i belongs to subgroup j, we let $\mathbf{x}_i = \mathbf{e}_j$, where $\mathbf{e}_j$ is the compositional vector with a 1 in the jth index. We now have an ANOVA-like model, but with a

compositional outcome.

This model has been considered in the literature where only the outcome is compositional, but previous solutions have either used an ILR transformation for $\mathbf{y}$ (Filzmoser, Hron, and Templ, 2018) or assumed that $\mathbf{y}|\mathbf{x}$ follows a Dirichlet distribution (Maier, 2014). Our model allows for a transformation-free and distribution-free solution for this problem. The formulation of our model in (3.4) shows that $\mathbf{B}_{j*} = E[\mathbf{y}|\mathbf{x} = \mathbf{e}_j]$, i.e., the rows of $\mathbf{B}$ simply interprets as the expectation for the j^{th} group. If we are interested in how $E[\mathbf{y}]$ changes between two groups j_1 and j_2, this change is represented by $\mathbf{B}_{j_1*} - \mathbf{B}_{j_2*}$. If the rows of $\mathbf{B}$ are all equal, this would indicate linear independence between $\mathbf{y}$ and $\mathbf{x}$.

3.3.2 Categorical outcome

We now $\mathbf{y}$ restrict to be categorical, meaning that each observation i belongs to one of $k = 1, \ldots, D_r$ groups. The standard model for this case would be a multinomial logistic model, using the ILR transformed $\mathbf{x}$ as covariates (Filzmoser, Hron, and Templ, 2018). However, we can use the model in (3.4), which allows for direct estimation of $E[y_k|\mathbf{x}] = P(\mathbf{y} = \mathbf{e}_k|\mathbf{x})$, $k = 1, \ldots, D_r$. This is equivalent to performing multinomial linear regression, with an identity link. The identity link is the canonical link here, as the covariates are compositional. Further restricting $\mathbf{x}$ to be categorical allows for interpretation of $\mathbf{B}_{j_1,k_1} - \mathbf{B}_{j_2,k_1}$ as $P(\mathbf{y} = \mathbf{e}_{k_1}|\mathbf{x} = \mathbf{e}_{j_1}) - P(\mathbf{y} = \mathbf{e}_{k_1}|\mathbf{x} = \mathbf{e}_{j_2})$, showing that our model is suitable for modeling risk differences between groups.

3.3.3 Discrete time series transition probabilities

A specific case of a categorical outcome and covariate is in estimating time-invariant transition probabilities for a first-order Markov process. An example of this class of problems is estimating the probability of firms or institutions transitioning between specific credit ratings (Jones, 2005). Observations may transition between $r = 1, \ldots, R$ states. In the ideal case, for each observation unit i, we observe their discrete state $\mathbf{y}_{i,t}$ over times $t = 0, \ldots, T$. We are then interested in estimating the probability that each observation moves to state j at time t, given that they are in state k at time $t-1$ (assuming transition probabilities are constant over time and between observation units). The interpretation of $\mathbf{B}$ from Sections 3.3.2 and 3.3.1 shows that if the covariate in (3.4), $\mathbf{y}_{i,t-1}$, and the outcome is, $\mathbf{y}_{i,t}$, then $\mathbf{B}_{jk} = P(\mathbf{y}_{i,t} = \mathbf{e}_j | \mathbf{y}_{i,t-1} = \mathbf{e}_k)$, which is exactly the transition probability we seek to estimate.

3.3.4 AR(1) model for compositional data

Rather than observing the states of each observation unit, we may only observe the percentage of observations in each state at each time. For example, Jones (2005) presents the case where for each year between 1984-2004, we only observe the percentage of commercial banks that belong to four different categories of credit quality. Our observed data is now the percentage of units in the different states at time t, $\mathbf{y}_t$. Specifically, y_{tj} is the percentage of observations belonging to state j at time t. Lee, Judge, and Zellner (1970), MacRae (1977), and Jones (2005) have shown that $E[\mathbf{y}_t | \mathbf{y}_{t-1}] = \mathbf{B}' \mathbf{y}_{t-1}$, where $\mathbf{B}_{ij}$ is again defined as $P(\mathbf{y}_{i,t} = \mathbf{e}_j | \mathbf{y}_{i,t-1} = \mathbf{e}_k)$. Thus, the direct regression

model in (3.5) can be used to estimate the individual transition probabilities, despite only observing aggregate data. For such settings, our model can be perceived as an AR(1) model for the compositional time series y_t.

3.4 Parameter Estimation

3.4.1 Generalized Method of Moments Approach

In order to estimate the entries of $\mathbf{B}$, we note that the model in (3.5) implies that

$$E[y_k|\mathbf{x}] = \sum_{j=1}^{D_s} B_{jk} x_j \, .$$

As we are only interested in the first moment of $\mathbf{y}|\mathbf{x}$, we use a generalized method of moments (GMM) (Hansen, 1982) approach and seek a function $\ell(\mathbf{B}; \mathbf{y}, \mathbf{x})$ such that

$$E_{\mathbf{B}_0}\left(\frac{d\ell}{d\mathbf{B}}\right) = 0 \, ,$$

where $\mathbf{B}_0$ is the true value of $\mathbf{B}$. A function ℓ which achieves this is, while also allowing for 0s in $\mathbf{y}_i$ and $\mathbf{x}_i$, the Kullback-Leibler distance (KLD) between two compositional vecotrs — the observed $\mathbf{y}_i$ and $E[\mathbf{y}_i|\mathbf{x}_i]$ (Fiksel et al., 2020), i.e.,

$$
\begin{aligned}
\ell &= \sum_{i=1}^{N} \mathrm{KLD}(y_i \parallel E[y_i \mid x_i]) \\
&= -\sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{E[y_{ik}|\mathbf{x}]}{y_{ik}}\right) \\
&= -\sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{\sum_{j=1}^{D_s} B_{jk}x_{ij}}{y_{ik}}\right) . \qquad (3.7)
\end{aligned}
$$

Letting $\mathcal{F} = \{\mathbf{B}; B_{jk} \geq 0, \sum_{k=1}^{D_r} \mathrm{B}_{jk} = 1\}$ be the constrained space for $\mathbf{B}$, minimizing (3.7) with respect to $\mathbf{B}$ is equivalent to maximizing the log-quasi-multinomial likelihood (Mullahy, 2015; Alenazi, 2019):

$$
\begin{aligned}
\min_{\mathbf{B}\in\mathcal{F}} \ell(\mathbf{B}; \mathbf{x}, \mathbf{y}) &= \min_{\mathbf{B}\in\mathcal{F}} -\sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{\sum_{j=1}^{D_s} B_{jk}x_{ij}}{y_{ik}}\right) \\
&= \max_{\mathbf{B}\in\mathcal{F}} \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\sum_{j=1}^{D_s} B_{jk}x_{ij}\right) \qquad (3.8)
\end{aligned}
$$

The multinomial quasi-likelihood belongs to the linear exponential family (Gourieroux, Monfort, and Trognon, 1984) and minimizing (3.7) (or equivalently, maximizing (3.8)) produces a consistent estimator for $\mathbf{B}_0$ (Gourieroux, Monfort, and Trognon, 1984; Papke and Wooldridge, 1996; Mullahy, 2015). In addition, Fiksel et al. (2020) show that (3.7) is convex with respect to $\mathbf{B}$, guaranteeing existence of a global minimum of (3.7).

3.4.2 An EM Algorithm for Maximizing the Objective Function

Alenazi (2019) also uses a GMM approach via minimization of the KLD between the observed and expected values for the compositional outcome in (3.3). Because the form of the conditional expected value in (3.3) is that used in multinomial logistic regression, the coefficients are unconstrained and Alenazi (2019) utilizes the Newton-Raphson (Böhning, 1992) algorithm for maximizing the log-quasi-multinomial likelihood. However, our model imposes constraints on the parameter space for $\mathbf{B}$ making it difficult to employ the Newton-Raphson algorithm to maximize (3.8).

We instead develop an EM algorithm for parameter estimation by maximization of (3.8) . We first present the algorithm for the special case where $\mathbf{y}_i$'s are categorical (Section 3.3.2). We introduce "missing" pseudo categories $\mathbf{x}_i^*$ such that $\mathbf{x}_i^*|\mathbf{x}_i \sim Multinomial(1, \mathbf{x}_i)$ and assume $\mathbf{y}_i|\mathbf{B}, x_{ij}^* = 1 \sim Multinomial(1, \mathbf{B}_{j*})$, thus using a proper likelihood for the outcome. We then arrive at the following likelihood of $\mathbf{y}|\mathbf{x}$ (marginalizing out the psuedo-categories $\mathbf{x}^*$):

$$p(\mathbf{y}|\mathbf{B},\mathbf{x}) = \prod_{i=1}^{N}\left(\sum_{j=1}^{D_s} p(x_{ij}^*=1)p(\mathbf{y}_i^*|\mathbf{B},x_{ij}^*=1)\right)$$

$$= \prod_{i=1}^{N}\left(\sum_{j=1}^{D_s} x_{ij}\prod_{k=1}^{D_r}(B_{jk})^{y_{ik}}\right)$$

$$= \prod_{i=1}^{N}\prod_{k=1}^{D_r}\left(\sum_{j=1}^{D_s} B_{jk}x_{ij}\right)^{y_{ik}} \tag{3.9}$$

Taking the log of (3.9) gives us the form of the objective function in (3.8). Letting $\mathrm{B}_{jk}^{(t)}$ denote the value of B_{jk} after iteration t, the expected complete log-likelihood becomes

$$Q(\mathbf{B}|\mathbf{B}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{D_2}\left[E[x_{ij}^*|x_{ij},y_{ik},\mathrm{B}_{jk}^{(t)}](log(x_{ij}) + \sum_{k=1}^{D_1} y_{ik}log(\mathrm{B}_{jk}))\right].$$

Noting that the M-step will require finding

$$\max_{\mathbf{B}\in\mathcal{F}}\sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} E[x_{ij}^*|x_{ij},y_{ik},\mathrm{B}_{jk}^{(t)}]y_{ik}\log(\mathrm{B}_{jk}) \tag{3.10}$$

we see that the terms in (3.10) for which $y_{ik} = 0$ will not influence the maximization. Thus, rather than evaluating both $E[x_{ij}^*|x_{ij},y_{ik}=0,\mathrm{B}_{jk}^{(t)}]$ and $E[x_{ij}^*|x_{ij},y_{ik}=1,\mathrm{B}_{jk}^{(t)}]$, we only have to evaluate the latter term. We thus introduce weights $\pi_{ijk}^{(t+1)}$ for the E-step at iteration $t+1$ which are equal to $E[x_{ij}^*|x_{ij},y_{ik}=1,\mathrm{B}_{jk}^{(t)}]$:

$$\pi_{ijk}^{(t+1)} = \frac{x_{ij}\mathrm{B}_{jk}^{(t)}}{\sum_{j=1}^{D_s} x_{ij}\mathrm{B}_{jk}^{(t)}},\ i = 1,\ldots,N, j = 1,\ldots,D_s, k = 1,\ldots,D_r$$

The expected complete log-likelihood is now

$$Q(\mathbf{B}|\mathbf{B}^{(t)}) = \sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik}\pi_{ijk}^{(t+1)}\log(\mathrm{B}_{jk}) \ ,$$

and the M-step from (3.10) becomes

$$\max_{\mathbf{B}\in\mathcal{F}} Q(\mathbf{B}|\mathbf{B}^{(t)}) = \max_{\mathbf{B}\in\mathcal{F}} \sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik}\pi_{ijk}^{(t+1)}\log(\mathrm{B}_{jk}) \ . \tag{3.11}$$

Due to the fact that $\sum_{k=1}^{D_r}\mathrm{B}_{jk} = 1$ for $j = 1,\ldots,D_s$, we can recognize the constrained maximization in (3.11) equivalent to maximizing $j = 1,\ldots,D_s$ weighted multinomial likelihoods. This implies the following M-step:

$$\mathrm{B}_{jk}^{(t+1)} = \frac{\sum_{i=1}^{N} y_{ik}\pi_{ijk}^{(t+1)}}{\sum_{k=1}^{D_r}\sum_{i=1}^{N} y_{ik}\pi_{ijk}^{(t+1)}},\ k = 1,\ldots,D_r,\ j = 1,\ldots,D_s \ .$$

Having developed an EM algorithm when we restrict the outcome $\mathbf{y}$ to be categorical, Theorem 1 now extends the EM algorithm to the general case when $\mathbf{y}$ is compositional:

Theorem 3. *Let* $f(t) = \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log\left(\sum_{j=1}^{D_s} B_{jk}^{(t)}x_{ij}\right)$ *be the value of the objective function after iteration* t *of the EM algorithm with compositional outcomes* $\mathbf{y}$*, using the same E and M steps as when* $\mathbf{y}$ *is categorical. Then* $f(t+1) - f(t) \geq 0$*, with strict inequality if* $Q(\mathbf{B}^{(t+1)}|\mathbf{B}^{(t)}) > Q(\mathbf{B}^{(t)}|\mathbf{B}^{(t)})$.

A proof is provided in the appendix. Theorem 1 allows use of the same EM

algorithm for estimation of $\mathbf{B}$, despite the fact that our approach is likelihood-free and only specifies $E[\mathbf{y}|\mathbf{x}]$.

3.5 A permutation test for linear independence

In the Chen, Zhang, and Li (2017) and Alenazi, 2019 models presented in (3.1) and (3.3), one can test whether each of the coefficients is equal to 0, using either bootstrapping (Efron and Tibshirani, 1994) or by estimating the standard errors of the coefficient estimates (Chen, Zhang, and Li, 2017; Mullahy, 2015). This is testing whether certain parts of $\mathbf{y}$ and $\mathbf{x}$ are associated with each other. We now present a permutation test for linear independence that can be applied to the direct regression method, and also can be adapted to the Chen, Zhang, and Li (2017) and Alenazi, 2019 models.

If $\mathbf{y}$ is linearly independent of $\mathbf{x}$, we have $E[\mathbf{y}|\mathbf{x}] = E[\mathbf{y}]$. The interpretation of our model in Section 3.3 shows that this is equivalent to restricting the model in (3.5) such that the rows of $\mathbf{B}$ are equal. We now develop a procedure for testing the following null hypothesis:

$$H_0 : E[\mathbf{y}] = \mathbf{B}_{1*} = \mathbf{B}_{2*} = \cdots = \mathbf{B}_{D_r*}$$

Letting $\boldsymbol{\mu} = E[\mathbf{y}]$, under the restricted model implied by H_0, the maximization task in (3.8) becomes

$$\max_{\boldsymbol{\mu} \in \mathbb{S}^{D_r}} \sum_{i=1}^{N} \sum_{k=1}^{D_r} y_{ik} \log(\mu_k) \, . \tag{3.12}$$

. The solution to the constrained maximization task in (3.12) leads to the

following estimate of $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}\mathbf{y}_i \,,$$

which is simply the arithmetic average of the observed $\mathbf{y}$. Letting $\bar{y}_k = \frac{1}{N}\sum_{i=1}^{N} y_{ik}$, under H_0 the log-quasi likelihood in (3.8) becomes

$$PLL_{H_0} = \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log(\bar{y}_k) \,.$$

Under the alternative hypothesis,

$$H_A : \mathbf{B}_{1*} \neq \mathbf{B}_{k*} \text{ for at least one value of } k \in \{2,\ldots,D_r\}$$

the log-quasi likelihood is that implied in (3.8):

$$PLL_{H_A} = \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik}\log\left(\sum_{j=1}^{D_s}\hat{B}_{jk}x_{ij}\right)$$

.

Comparing the log-quasi likelihoods under H_0 and H_A leads to the following test statistic of interest:

$$\lambda = PPL_{H_A} - PPL_{H_0}$$

which is equivalent to the log-quasi likelihood ratio between the restricted and full models. To obtain the distribution of λ under H_0, we use the following Monte Carlo permutation testing procedure (Good, 2005):

Step 1: Obtain λ^{obs} using the observed $\mathbf{x}$ and $\mathbf{y}$.

Step 2: Randomly permute the observed $\mathbf{x}$ to break any dependence between $\mathbf{x}$ and $\mathbf{y}$

Step 3: Obtain λ^{perm} using the permuted $\mathbf{x}$ and observed $\mathbf{y}$.

Step 4: Repeat Steps 2-3 $b = 1, \ldots, B$ times, obtaining λ^{perm_b} for each permutation. In practice, setting $B = 1000$ appears to give good precision (Zeng et al., 2015).

Step 5: Calculate the p-value, $p = \frac{1}{B} \sum_{b=1}^{B} I(\lambda^{perm_b} \geq \lambda^{obs})$

The permutation test procedure allows for testing of whether changing any part of the compositional $\mathbf{x}$ is associated with a linear change in the expected value of $\mathbf{y}$. Furthermore, this procedure can be adopted for use in the models presented by Chen, Zhang, and Li (2017) and Alenazi (2019), either using the normal likelihood for the ILR transformed outcome, or the log-quasi likelihood using the conditional expected value formulation in (3.3). The permutation test procedure can also be modified for testing the null hypothesis that rows $\mathbf{B}_{j_1}$ and $\mathbf{B}_{j_2}$ are equal. As discussed in Section 3.3, if two rows of $\mathbf{B}$ are equal, $E[\mathbf{y}|\mathbf{x}]$ reduces to (3.6). We can use this expectation to obtain the log-quasi likelihood under the null hypothesis that $\mathbf{B}_{j_1}$ and $\mathbf{B}_{j_2}$ are equal.

3.6 Simulation studies

3.6.1 Model comparison study

We first perform simulations to compare the performance of the direct regession model with that of the Chen, Zhang, and Li (2017) model and the Alenazi (2019) model across situations when only one of the three models is correctly specified. To generate realistic data, we first fit each model to two datasets with a compositional outcome and explanatory variable: the Education dataset

(Section 3.7.1) and the White Cells dataset (Section 3.7.2). For the Alenazi (2019) model, we let $t(\mathbf{x}) = ilr(\mathbf{x})$. These fitted coefficients are then used as the true coefficient values for each model when simulating data. Compositional covariates $\mathbf{x}_i$ $(i = 1, \ldots, N; N = 100, 250, 500, 1000)$ were simulated independently such that $x_i \sim Dirichlet(1, 1, 1)$. Because our direct regression model and the Alenazi (2019) model both directly specify $E[\mathbf{y}_i|\mathbf{x}_i]$, we used the coefficients for each model from the two datasets to obtain the true conditional expected values, and then simulated $\mathbf{y}_i|\mathbf{x}_i \sim Dirichlet(10 \cdot E[\mathbf{y}_i|\mathbf{x}_i])$ for each model. For the Chen, Zhang, and Li (2017) model, we simulated $ilr(\mathbf{y}_i)|\mathbf{x}_i \sim \mathcal{N}(E[ilr(\mathbf{y}_i)|\mathbf{x}_i], 1)$, and used $\mathbf{y}_i = ilr^{-1}(\mathbf{y}_i)$ as the compositional outcome.

Each of the three models were fit on the simulated data. To compare models, we generated a large, independent test set and obtained the true $E[\mathbf{y}_i|\mathbf{x}_i]$ for each observation. We then obtain the average KLD between the true and estimated conditional means in this independent set. This full process is repeated 10,000 times for every combination of N, true data generating mechanism, and dataset.

For ease of comparison, Figure 3.2 shows the log KLD for each simulation setting, averaged across all 10,000 simulations. Unsurprisingly, the correctly specified model performs the best in conditional mean estimation across almost all settings. Interestingly, the Chen, Zhang, and Li (2017) model appears to perform much worse when it is misspecified, as compared to the direct regression model and the Alenazi (2019) model. Overall, these results show that each of these models can be used to model compositional regression models,

and that the KLD (either estimated on a test set or through cross-validation) is a valid metric for model comparison.

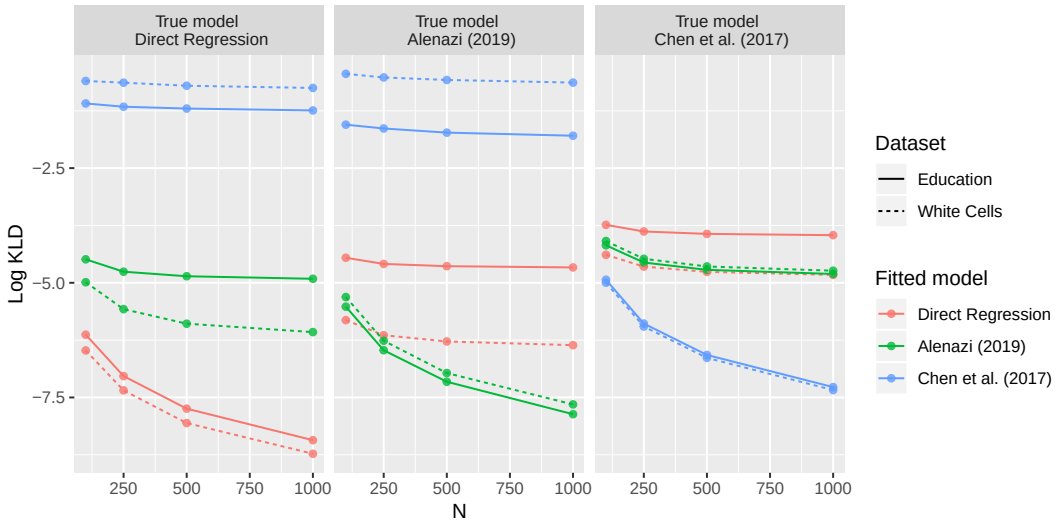


Figure 3.2: Log KLD estimated using a test set, across various sample sizes and true models. Each column represents a different true model for the compositional outcome, with two true coefficients values estimated on different datasets (solid and dashed lines). Each color shows the estimated Log KLD based on the fitted model.

3.6.2 Direct regression on different data generating mechanisms

Because the direct regression model does not specify a likelihood for $\mathbf{y}|\mathbf{x}$, we compare performance of the direct regression model across different data generating mechanisms that share the same conditional mean model. As in Section (3.6.1), we estimate the coefficients of the direct regression model on the same two datasets, and generate covariates $\mathbf{x}_i$ using a uniform Dirichlet distribution. We then generated $\mathbf{y}_i|\mathbf{x}_i$ using three data generating mechanisms presented by Murteira and Ramalho (2016):

1. Dirichlet: The compositional outcome $\mathbf{y}_i$ is directly generated via the

model $\mathbf{y}_i|\mathbf{x}_i \sim Dirichlet(10 \cdot \mathbf{B}^{'}\mathbf{x}_i)$

2. Multinomial (proportion): We first generate an individual "sample-size" n_i from a $Discrete - Uniform(1, 30)$ distribution. Individual counts are generated via $\mathbf{y}_i^*|\mathbf{x}_i \sim Multinomial(n_i, \mathbf{B}^{'}\mathbf{x}_i)$, and the compositional outcome $\mathbf{y}_i$ is defined such that $y_{ik} = \frac{y_{ik}^*}{\sum_{k=1}^{3} y_{ik}^*}$

3. Dirichlet-multinomial (proportion): We introduce over-dispersion into the multinomial data generating scheme, by first simulating $\mathbf{p}_i|\mathbf{x}_i \sim Dirichlet(10 \cdot \mathbf{B}^{'}\mathbf{x}_i)$. Rather than simulating $\mathbf{y}_i^*|\mathbf{x}_i \sim Multinomial(n_i, \mathbf{B}^{'}\mathbf{x}_i)$, we instead simulate $\mathbf{y}_i^*|\mathbf{x}_i \sim Multinomial(n_i, \mathbf{p}_i)$. The compositional outcome $\mathbf{y}_i$ is again defined such that $y_{ik} = \frac{y_{ik}^*}{\sum_{k=1}^{3} y_{ik}^*}$

The fitted direct regression models are evaluated via KLD on a test set, as in Section (3.6.1). Figure 2 shows that while the (log) KLD is similar across all data generating mechanisms, the model performs slightly worse for the models with higher variance for the compositional outcome. However, when the other two (incorrectly specified) models are fit to this simulated data, the direct regression model outperforms these models across all data generating mechanisms (Appendix Figure 3.7), again showing the importance of correctly specifying the conditional mean for the compositional outcome.

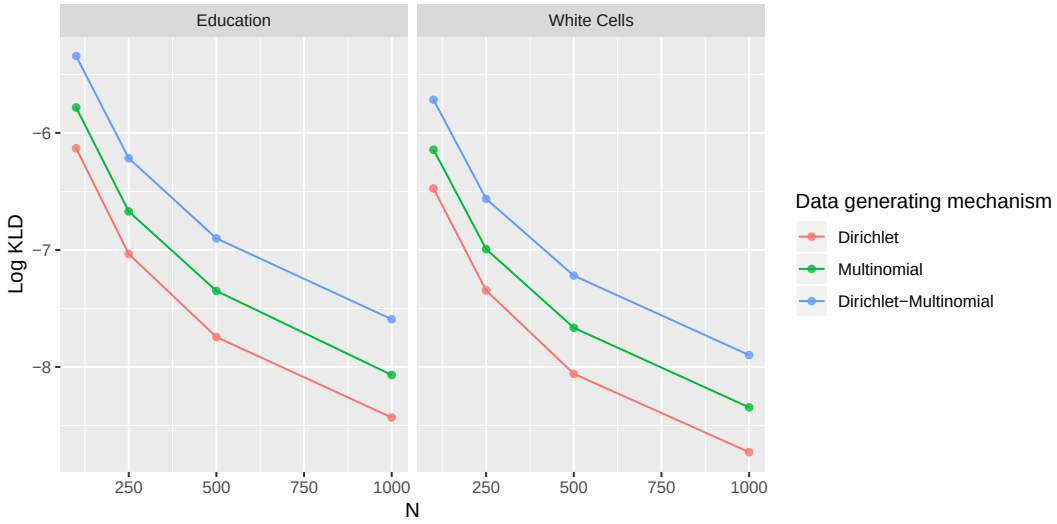


Figure 3.3: KLD estimated using a test set, across various sample sizes and data generating mechanisms, with the conditional mean specified via the direct regression model. Each column represents a different true value for **B**, based on the two different real-world datasets. Each color shows the estimated KLD for different data generating mechanisms for the compositional outcome.

3.6.3 Evaluating the Type-I and Type-II error rates of the global linear independence test

To evaluate the testing procedure introduced in Section 3.5 in terms of Type-I and Type-II error rates, we perform a simulation study that we detail in the Appendix. In summary, we observe that when **y** is linearly independent of **x**, our procedure produces pre-specified Type-I error rates, regardless of the data generating mechanism for **y**. We also observe that the permutation test generally has high power to detect linear relationships between E[**y**] and **x**, although this is not the case for smaller sample sizes (n=100) when the linear relationship is weak. Finally, we observe that when the true conditional mean is that specified by the direct regression model, but the model is specified via

Chen, Zhang, and Li (2017) model, the permutation test has lower power to detect dependence between $E[\mathbf{y}]$ and $\mathbf{x}$.

3.7 Applications

To show that our method can realistically use data to address scientific questions in an interpretable manner, we now apply our method to two datasets which have a compositional predictor and a compositional outcome.

3.7.1 Educational status of mothers and fathers in European countries

Parental educational attainment has a large effect on child outcomes (Dubow, Boxer, and Huesmann, 2009). Filzmoser, Hron, and Templ (2018) provide a dataset that contains the percent of fathers and mothers with low, medium, and high education levels in 31 European countries. The question of interest is how the percentage of fathers with a given education level relate to the percentage of mothers with different education levels, across the 31 countries. We let y_{ik} be the percentage of fathers with education level k (1 = low (pre-primary, primary or lower secondary education), 2 = medium (upper secondary edu 3 = high (first stage of tertiary education and second stage of tertiary education)) (Eurostat, 2015) in country i, and x_{ij} be the percentage of mothers with education level j.

Fitting the model in (3.5) leads to the following estimate of $\mathbf{B}$:

$$\hat{\mathbf{B}} = \begin{pmatrix} .91 & .05 & .04 \\ .00 & .91 & .09 \\ .00 & .14 & .86 \end{pmatrix}$$

which shows high correlation between the educational attainment status of fathers and mothers (independence test p-value=0). The coefficients and 95% confidence regions, obtained via bootstrap, are shown in Figure 3.4. There is noticeably more uncertainty in estimation of $\mathbf{B}_{3*}$ than in the other rows of $\mathbf{B}$. In addition, there is very little uncertainty in $\hat{\mathbf{B}}_{2,1}$

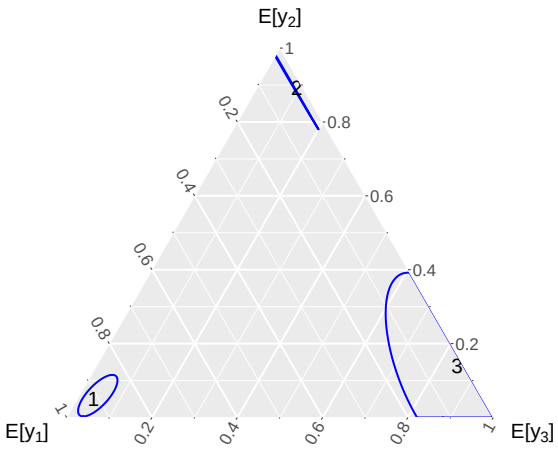


Figure 3.4: Visualization of the coefficients for regression the percentage of fathers of a given education level on the percentage of mothers of a given education level. Each row of $\hat{\mathbf{B}}$ is labeled with a number in the ternary diagram. The 95% confidence region for each row is drawn in blue.

The analytical interpretation of $\hat{\mathbf{B}}$ means that increasing the percentage of mothers with a medium level of education level by .10, while decreasing the percentage of mothers with a low level of education level by .10, is associated with a change in the percentage of fathers with low, medium, and high educational status of -.09, .09, and .01, respectively. Similar affects are seen for other

changes of the percentage of mothers with a given educational status.

To visualize the model fit, we first obtain predicted values for each of the father educational compositions, using leave-one-out cross-validation (LOOCV) (Friedman, Hastie, and Tibshirani, 2001), based off the mother educational compositions in each country. Figure 3.5 shows the observed versus predicted percentage of fathers with each level of education, across the 31 countries. The predicted percentages are all very close to the observed percentages, showing that our simple model is not only interpretable, but also appears to fit the observed data well.

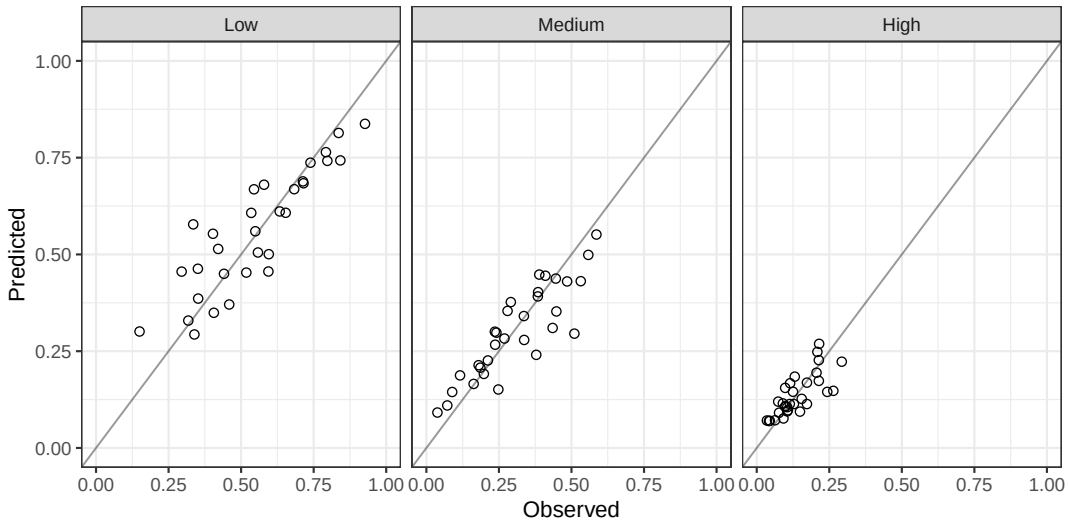


Figure 3.5: Observed versus predicted father educational attainment compositions across each of the 31 countries. The grey line represents the identity line.

We also compare our model to the models presented by Chen, Zhang, and Li (2017) and Alenazi (2019) using the KLD between the observed $\mathbf{y}$ and predicted $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is estimated via LOOCV for all three methods. Each of the three methods had a KLD of .024, indicating similar model fit.

3.7.2 White cell composition analysis

Aitchison (2003) and Alenazi (2019) consider a dataset in which the proportions of white blood cell types (granulocytes, lymphocytes, and monocytes) in 30 blood samples are determined by both a time-consuming microscopic analysis and an automated image analysis. The microscopic analysis is known to produce accurate results, while the accuracy of the image analysis is unknown. If the estimated compositions from the microscopic analysis can be predicted by the compositions estimated by the image analysis, it would be time-saving to use the automated image analysis in the future.

We let y_{ik} and x_{ij} be the estimated composition of white blood cell type k and j (1 = granulocytes, 2 = lymphocytes, 3 = monocytes) by the microscopic and image analysis, respectively. The estimate of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \begin{pmatrix} .97 & .03 & .00 \\ .00 & 1.00 & .00 \\ .00 & .04 & .96 \end{pmatrix}$$

which again shows high correlation between the compositional outcome and explanatory variables (independence test p-value=0). Because $\hat{\mathbf{B}}$ is extremely close to the identity matrix (i.e. perfect correlation), visualization of $\hat{\mathbf{B}}$ provides little additional benefit in interpretation and we do not plot $\hat{\mathbf{B}}$ in a ternary diagram.

If the image analysis estimates a white blood cell composition of $(.65, .26, .09)$, the average estimated composition for the image analysis, we would predict that the microscopic analysis estimates a composition of $(.63, .28, .09)$. Increasing one part of the estimated composition from the image analysis would also

lead to a very similar predicted increase in the estimated composition from the microscopic analysis. Figure 3.6 again shows that our method produces extremely accurate predictions, obtained via LOOCV.

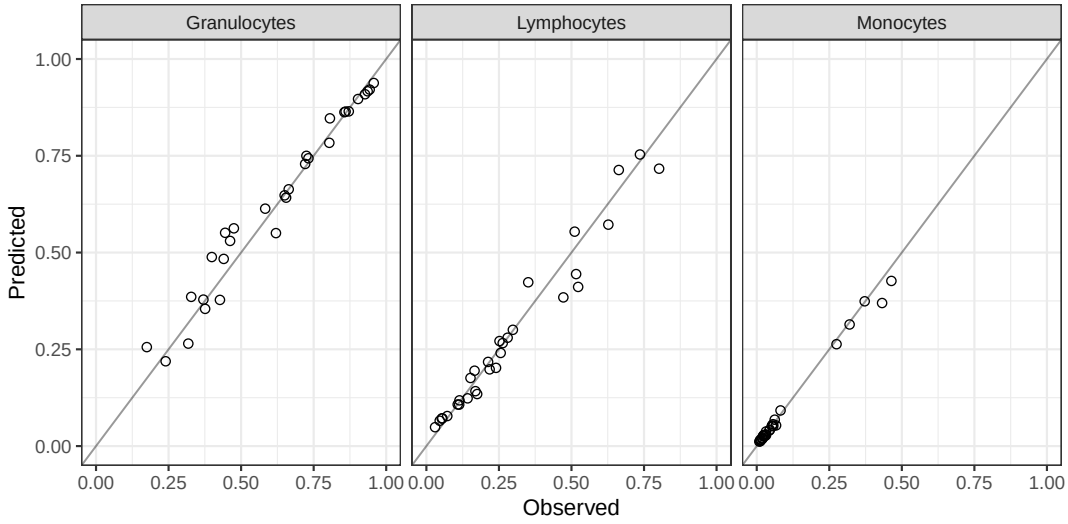


Figure 3.6: Observed versus predicted white blood cell composition estimates using the microscopic analysis from each of the 30 samples. The grey line represents the identity line.

Finally, we again compare our method to the methods presented in Section (3.2) using the KLD. As in the analysis in (3.7.1), the models perform nearly identically, with direct regression model and the model from Chen, Zhang, and Li (2017) producing a KLD of .005, and the model from Alenazi (2019) producing a KLD of .006. These two analyses show that our method is not only more interpretable, but also comes without loss of fidelity to the observed data.

3.8 Discussion

In this manuscript, we have introduced a novel direct regression model for compositional outcomes and explanatory variables. This direct regression model offers a simple interpretation of the regression coefficients, as opposed to currently used transformation-based methods. The simple interpretation of the direct regression model's coefficients facilitates the use of this model by practitioners who are not deeply familiar with compositional data transformations, without having to resort to graphical techniques for visualizing the response surface. In addition to its simplicity, the direct regression model can accurately approximate observed scientific data, as shown in Sections 3.7.1 and 3.7.2. Fast parameter estimation is obtained through a likelihood-free EM algorithm, and a global null hypothesis test is performed via a quasi-likelihood ratio test.

One important future direction is developing a robust workflow for model comparison and selection for compositional regression problems. Although we have shown the potential of comparing the estimated KLD between models, there may be additional graphical and analytical tools that may yield better insight. Another important future direction is extending the direct regression model to allow for either continuous covariates or multiple compositional covariates, while maintaining simple interpretations for the compositional covariate coefficients. Current models for this problem simply extend the Chen, Zhang, and Li (2017) model by including the continuous covariates in the model Morais, Thomas-Agnan, and Simioni, 2018. A potential solution is to use the direct regression model to model the partial dependence (Greenwell,

2017) between the compositional outcome and the compositional covariates of interest, but we leave this for future work.

3.9 Appendix

3.9.1 Proofs

Proof of Theorem 1. We adopt this proof from the proof of Theorem 2.1 in Yao (2013). For $i = 1, \ldots, N$ and $k = 1, \ldots, D_1$, let $z_{ik}^{(t+1)}$ be a discrete random variable such that

$$P\left(z_{ik}^{(t+1)} = \frac{\mathrm{B}_{jk}^{(t+1)}}{\mathrm{B}_{jk}^{(t)}}\right) = \frac{x_{ij}\mathrm{B}_{jk}^{(t)}}{\sum_{j=1}^{D_s} x_{ij}\mathrm{B}_{jk}^{(t)}} = \pi_{ijk}^{(t+1)}, j = 1, \ldots, D_s$$

We then have

$$
\begin{aligned}
f(\mathrm{B}^{(t+1)}) - f(\mathrm{B}^{(t)}) &= \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\frac{\sum_{j=1}^{D_s} B_{jk}^{(t+1)} x_{ij}}{\sum_{j=1}^{D_s} B_{jk}^{(t)} x_{ij}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\sum_{j=1}^{D_s} \frac{B_{jk}^{(t)} x_{ij}}{\sum_{j=1}^{D_s} B_{jk}^{(t)} x_{ij}} \cdot \frac{B_{jk}^{(t+1)} x_{ij}}{B_{jk}^{(t)} x_{ij}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(\sum_{j=1}^{D_s} \pi_{ijk}^{(t+1)} \cdot \frac{B_{jk}^{(t+1)} x_{ij}}{B_{jk}^{(t)} x_{ij}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \log\left(E[z_{ik}^{(t+1)}]\right) \\
&\geq \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} E[\log\left(z_{ik}^{(t+1)}\right)] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{D_r} y_{ik} \sum_{j=1}^{D_s} \pi_{ijk}^{(t+1)} log\left(\frac{\mathrm{B}_{jk}^{(t+1)}}{\mathrm{B}_{jk}^{(t)}}\right) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik} \pi_{ijk}^{(t+1)} \left[\log\left(\mathrm{B}_{jk}^{(t+1)}\right) - \log\left(\mathrm{B}_{jk}^{(t)}\right)\right]
\end{aligned}
$$

Because the M-step in (3.11) is the same regardless of whether $\mathbf{y}$ is categorical or compositional, this implies that

$$
\sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik} \pi_{ijk}^{(t+1)} \log\left(\mathrm{B}_{jk}^{(t+1)}\right) \geq \sum_{i=1}^{N}\sum_{k=1}^{D_r}\sum_{j=1}^{D_s} y_{ik} \pi_{ijk}^{(t+1)} \log\left(\mathrm{B}_{jk}^{(t)}\right)
$$

we have $f(\mathrm{B}^{(t+1)}) - f(\mathrm{B}^{(t)}) \geq 0$, with $f(\mathrm{B}^{(t+1)}) - f(\mathrm{B}^{(t)}) > 0$ if $Q(\mathbf{B}^{(t+1)}|\mathbf{B}^{(t)}) > Q(\mathbf{B}^{(t)}|\mathbf{B}^{(t)})$.

3.9.2 Additional Figures

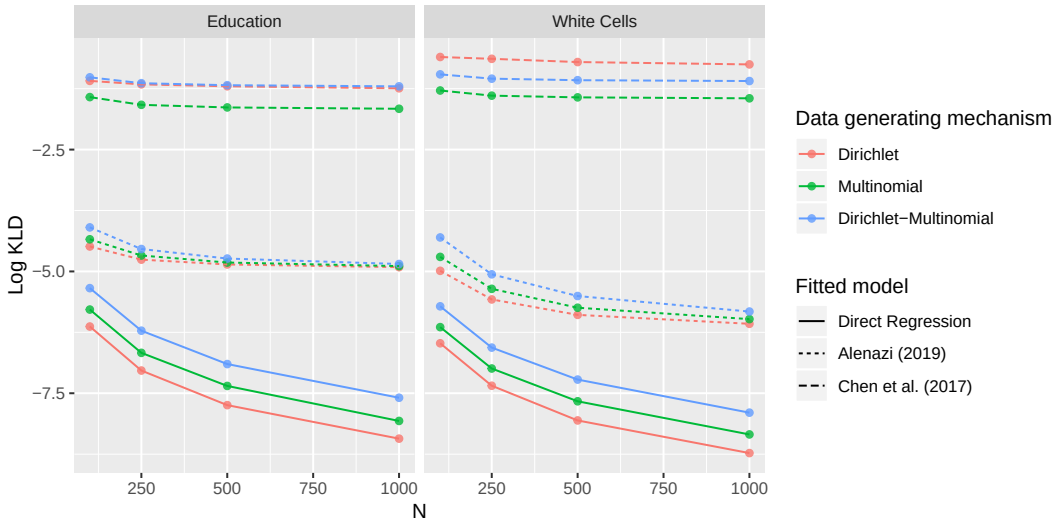


Figure 3.7: Comparison of models via Log KLD, when the direct regression model specification is the correct conditional mean. The correctly specified direct regression model outperforms the other two models, across data generating mechanisms, coefficient values, and sample sizes.

3.9.3 Coefficient values for the ILR regression model

Model 1

$$\beta_{01} = 1,\ \beta_{11} = 2,\ \beta_{21} = -1$$

$$\beta_{02} = -2,\ \beta_{12} = -1,\ \beta_{22} = 2$$

Model 2

$$\beta_{01} = 1,\ \beta_{11} = .333,\ \beta_{21} = -.333$$

$$\beta_{02} = -2,\ \beta_{12} = -.333,\ \beta_{22} = .333$$

Model 3

$$\beta_{01} = 1,\ \beta_{11} = 2,\ \beta_{21} = 0$$

$$\beta_{02} = -2,\ \beta_{12} = -1,\ \beta_{22} = 0$$

3.9.4 Simulation study to evaluate Type-I and Type-II error rates for the global independence test

We again generated $\mathbf{x}_i$ independently from a uniform Dirichlet distribution for $i = 1, \ldots, N$, with $N = 100, 250, 500, 1000$. We then generated $\mathbf{y}_i|\mathbf{x}_i$ using the three data generating mechanisms introduced in Section 3.6.2.

To evaluate the Type-I error rate, we generated data via the direct regression model by setting each row of $\mathbf{B}$ to be $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, which implies that $E[\mathbf{y}|\mathbf{x}] = E[\mathbf{y}] = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. We then simulated 10,000 data sets for each combination of of the 3 data generating mechanisms and 4 sample sizes. Table 2 shows the percentage of the simulations where the observed p-value was below .05. Across all the sample sizes and data generating mechanisms for $\mathbf{y}$, we see that all observed Type-I error rates are very close to the nominal .05 rate, showing that the permutation test is well calibrated.

True Distribution	N=100	N=250	N=500	N=1000
Dirichlet	.050	.052	.051	.052
Multinomial	.054	.050	.052	.047
Dirichlet-Multinomial	.050	.050	.048	.052

Table 3.2: Empirical Type-I error rates across different sample sizes and data generating distributions for $\mathbf{y}$.

For evaluating the Type-II error rate when the direct regression model is correctly specified, we used the three different values for $\mathbf{B}$:

$$\mathbf{B}^{(1)} = \begin{pmatrix} .90 & .05 & .05 \\ .05 & .90 & .05 \\ .05 & .05 & .90 \end{pmatrix}; \mathbf{B}^{(2)} = \begin{pmatrix} .40 & .30 & .30 \\ .30 & .40 & .30 \\ .30 & .30 & .40 \end{pmatrix}; \mathbf{B}^{(3)} = \begin{pmatrix} .90 & .05 & .05 \\ .33 & .33 & .33 \\ .33 & .33 & .33 \end{pmatrix}$$

The interpretations of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ were introduced in Section 3.3. $\mathbf{B}^{(3)}$ represents the setting when y_1 and x_1 are highly correlated, but increasing x_2 at the expense of x_3 (and vice-versa) do not lead to any changes in $E[\mathbf{y}]$.

Table 3 shows the percentage of simulations for each setting where the observed p-value was greater than .05. For $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(3)}$, the permutation test shows extremely good performance in terms of Type-II error. Because the rows of $\mathbf{B}^{(2)}$ are fairly close to being equal, the method unsurprisingly has a high Type-II error rate for $N = 100$. Interestingly, the Type-II error rates differ across the three data generating mechanisms. As N increases, the Type-II error rate decreases across all data generating mechanisms, with a Type-II error rate close to 0 when $N = 1000$.

Value for $\mathbf{B}$	True Distribution	N=100	N=250	N=500	N=1000
$\mathbf{B}^{(1)}$	Dirichlet	.000	.000	.000	.000
	Multinomial	.000	.000	.000	.000
	Dirichlet-Multinomial	.000	.000	.000	.000
$\mathbf{B}^{(2)}$	Dirichlet	.582	.152	.006	.000
	Multinomial	.696	.322	.049	.001
	Dirichlet-Multinomial	.812	.549	.211	.018
$\mathbf{B}^{(3)}$	Dirichlet	.000	.000	.000	.000
	Multinomial	.000	.000	.000	.000
	Dirichlet-Multinomial	.003	.000	.000	.000

Table 3.3: Type-II error rates for the direct regression model across different values of $\mathbf{B}$, data generating mechanisms, and sample sizes.

We also evaluated the Type-II error rate of our method when the true

model is the Chen, Zhang, and Li (2017) model. We specify $E[ilr(\mathbf{y}_i)_k]$ via the model in (3.1) and provide coefficient values in the appendix. Outcomes $\mathbf{y}_i$ were generated by first simulating $ilr(\mathbf{y}_i)_k \sim \mathcal{N}(E[ilr(\mathbf{y}_i)_k|\mathbf{x}_i], 1)$ and then setting $\mathbf{y}_i = ilr^{-1}(ilr(\mathbf{y}_i))$. The permutation test achieved a Type-II error rate of 0 for all sample sizes and coefficient values, showing robustness to incorrect specification.

Finally, we evaluate the Type-II error rate of a likelihood ratio permutation test using the Chen, Zhang, and Li (2017) model. We use a normal likelihood for the ILR transformed outcomes, and estimate the coefficients and standard errors via maximum likelihood, as in Chen, Zhang, and Li (2017). When the ILR model is correctly specified, using the coefficient values specified in the appendix, the Type-II error rate is 0 across all sample sizes. However, when the true conditional mean is that specified by the direct regression model, comparing Table 4 to Table 3shows the ILR regression model to have lower power than the direct regression model.

Value for $\mathbf{B}$	True Distribution	N=100	N=250	N=500	N=1000
$\mathbf{B}^{(1)}$	Dirichlet	.000	.000	.000	.000
	Multinomial	.000	.000	.000	.000
	Dirichlet-Multinomial	.000	.000	.000	.000
$\mathbf{B}^{(2)}$	Dirichlet	.642	.225	.017	.000
	Multinomial	.914	.854	.743	.515
	Dirichlet-Multinomial	.909	.834	.692	.405
$\mathbf{B}^{(3)}$	Dirichlet	.000	.000	.000	.000
	Multinomial	.320	.014	.000	.000
	Dirichlet-Multinomial	.231	.004	.000	.000

Table 3.4: Type-II error rates for the Chen, Zhang, and Li (2017) model, using different values of $\mathbf{B}$, data generating mechanisms, and sample sizes.

References

Mullahy, John (2015). "Multivariate fractional regression estimation of econometric share models". In: *Journal of Econometric Methods* 4.1, pp. 71–100.

Murteira, José MR and Joaquim JS Ramalho (2016). "Regression analysis of multivariate fractional data". In: *Econometric Reviews* 35.4, pp. 515–552.

Papke, Leslie E and Jeffrey M Wooldridge (1996). "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates". In: *Journal of applied econometrics* 11.6, pp. 619–632.

Templ, Matthias, Peter Filzmoser, and Clemens Reimann (2008). "Cluster analysis applied to regional geochemical data: problems and possibilities". In: *Applied Geochemistry* 23.8, pp. 2198–2213.

Dumuid, Dorothea, Tyman E Stanford, Josep-Antoni Martin-Fernández, Željko Pedišić, Carol A Maher, Lucy K Lewis, Karel Hron, Peter T Katzmarzyk, Jean-Philippe Chaput, Mikael Fogelholm, et al. (2018). "Compositional data analysis for physical activity, sedentary time and sleep research". In: *Statistical methods in medical research* 27.12, pp. 3726–3738.

Lin, Wei, Pixu Shi, Rui Feng, and Hongzhe Li (2014). "Variable selection in regression with compositional covariates". In: *Biometrika* 101.4, pp. 785–797.

Leite, Maria Léa Corrêa (2016). "Applying compositional data methodology to nutritional epidemiology". In: *Statistical methods in medical research* 25.6, pp. 3057–3065.

Hron, Karel, Peter Filzmoser, and Katherine Thompson (2012). "Linear regression with compositional explanatory variables". In: *Journal of Applied Statistics* 39.5, pp. 1115–1128.

McGregor, DE, J Palarea-Albaladejo, PM Dall, K Hron, and SFM Chastin (2019). "Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns". In: *Statistical methods in medical research*, p. 0962280219864125.

Egozcue, Juan José, Josep Daunis-I-Estadella, Vera Pawlowsky-Glahn, Karel Hron, and Peter Filzmoser (2012). "Simplicial regression. The normal model". In: *Journal of Applied Probability and Statistics* 6.1 & 2, pp. 87–108.

Hijazi, Rafiq H and Robert W Jernigan (2009). "Modelling compositional data using Dirichlet regression models". In: *Journal of Applied Probability & Statistics* 4.1, pp. 77–91.

Wang, Huiwen, Liying Shangguan, Junjie Wu, and Rong Guan (2013). "Multiple linear regression modeling for compositional data". In: *Neurocomputing* 122, pp. 490–500.

Chen, Jiajia, Xiaoqin Zhang, and Shengjia Li (2017). "Multiple linear regression with compositional response and covariates". In: *Journal of Applied Statistics* 44.12, pp. 2270–2285.

Alenazi, Abdulaziz (2019). "Regression for Compositional Data With Compositional Data as Predictor Variables With or Without Zero Values". In: *Journal of Data Science* 17.1, pp. 219–237.

Filzmoser, Peter, Karel Hron, and Matthias Templ (2018). *Applied Compositional Data Analysis With Worked Examples in R*. Cham, Switzerland: Springer.

Aitchison, John (1986). *The Statistical Analysis of Compositional Data*. London: Chapman & Hall,

Morais, Joanna, Christine Thomas-Agnan, and Michel Simioni (2018). "Interpretation of explanatory variables impacts in compositional regression models". In: *Austrian Journal of Statistics* 47.5, pp. 1–25.

Egozcue, Juan José, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal (2003). "Isometric logratio transformations for compositional data analysis". In: *Mathematical Geology* 35.3, pp. 279–300.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Nguyen, Thi Huong An, Thibault Laurent, Christine Thomas-Agnan, and Anne Ruiz-Gazen (2018). "Analyzing the impacts of socio-economic factors on French departmental elections with CODA methods". In:

Tsagris, Michail (2015). "Regression analysis with compositional data containing zero values". In: *arXiv preprint arXiv:1508.01913*.

Hamilton, Nicholas E. and Michael Ferry (2018). "ggtern: Ternary Diagrams Using ggplot2". In: *Journal of Statistical Software, Code Snippets* 87.3, pp. 1–17. DOI: 10.18637/jss.v087.c03.

Maier, Marco J (2014). "DirichletReg: Dirichlet regression for compositional data in R". In:

Jones, Mr Matthew T (2005). *Estimating Markov transition matrices using proportions data: an application to credit risk*. 5-219. International Monetary Fund.

Lee, Tsoung-Chao, George G Judge, and Arnold Zellner (1970). "Estimating the parameters of the Markov probability model from aggregate time series data". In:

MacRae, Elizabeth Chase (1977). "Estimation of time-varying Markov processes with aggregate data". In: *Econometrica: journal of the Econometric Society*, pp. 183–198.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Fiksel, Jacob, Abhirup Datta, Agbessi Amouzou, and Scott Zeger (2020). "Generalized Bayesian Quantification Learning". In: *arXiv e-prints*, arXiv:2001.05360, arXiv:2001.05360. arXiv: 2001.05360 [stat.ME].

Gourieroux, Christian, Alain Monfort, and Alain Trognon (1984). "Pseudo maximum likelihood methods: Theory". In: *Econometrica: journal of the Econometric Society*, pp. 681–700.

Böhning, Dankmar (1992). "Multinomial logistic regression algorithm". In: *Annals of the institute of Statistical Mathematics* 44.1, pp. 197–200.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Good, Phillip (2005). *Permutation, parametric, and bootstrap tests of hypotheses*. 3rd. Springer.

Zeng, Ping, Yang Zhao, Hongliang Li, Ting Wang, and Feng Chen (2015). "Permutation-based variance component test in generalized linear mixed model with application to multilocus genetic association study". In: *BMC medical research methodology* 15.1, p. 37.

Dubow, Eric F, Paul Boxer, and L Rowell Huesmann (2009). "Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations". In: *Merrill-Palmer quarterly (Wayne State University. Press)* 55.3, p. 224.

Eurostat (2015). *Archive:Living condition statistics - family situation of today's adults as children*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Living_condition_statistics_-_family_situation_of_today\%27s_adults_as_children\&oldid=231142#Parents.E2.80.99_level_of_education.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Aitchison, John (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press.

Greenwell, Brandon M (2017). "pdp: an R Package for constructing partial dependence plots". In: *The R Journal* 9.1, pp. 421–436.

Yao, Weixin (2013). "A note on EM algorithm for mixture models". In: *Statistics & Probability Letters* 83.2, pp. 519–526.

Chapter 4

Generalized Bayesian Quantification Learning for Dataset Shift

4.1 Introduction

Classifiers are often developed with the goal of obtaining accurate predictions for individual units. For example, risk prediction models have been developed for identifying patients at high risk of cardiovascular disease. The outputs from these classifiers are used to guide decision making for individuals in the clinical setting (Moons et al., 2012). However, in some applications, the objective is not individual level predictions, but rather to learn about population-level distributions of a given outcome. Examples include sentiment analysis for Twitter users (Giachanou and Crestani, 2016), estimating the prevalence of chronic fatigue syndrome (Valdez et al., 2018), and cause of death distribution estimation from verbal autopsies (King, Lu, et al., 2008; McCormick et al., 2016; Serina et al., 2015; Byass et al., 2012; Miasnikof et al., 2015).

The task of predicting the population distribution of unobserved true outcomes (labels) based on observed covariates has been termed *quantification* (Forman, 2005; Bella et al., 2010; González et al., 2017; Pérez-Gállego et al., 2019) in the machine learning literature. Since the covariates are usually passed through a classifier to obtain predicted labels, quantification can be viewed as prevalence estimation using these predicted labels. Quantification requires building a classifier which can predict an outcome y using variables $\mathbf{x}$. This can be done by obtaining training data with observed outcomes y and variables $\mathbf{x}$ that can be used to train a classifier, or alternatively, creating a classifier based on expert knowledge (Kalter et al., 2015). In either case, the classifier is then used to predict labels in the test set representing the population of interest where we want to estimate the distribution of the categorical outcome y, but only observe $\mathbf{x}$. The predicted classes (or probabilities) for individuals in the test set are then aggregated to obtain an estimate of the distribution of the outcome in this population, $p_{test}(y)$ (Forman, 2005).

Quantification is distinct from building a classifier. It also goes beyond the task of training a classifier to accurately predict individual labels as common methods for quantification adjust output from inaccurate classifiers to improve quantification (Forman, 2008; Bella et al., 2010). However, these adjustments currently rely on estimating the classifier's true and false positive rates (or their multi-class equivalents) from the training data and assumes that these rates are the same in the test set. A review of the current approaches to quantification is provided in Section 4.2. This is similar to approaches used for *transportability* of clinical trial results, which use a weighted average of covariate conditional

treatment effects obtained from the study sample to estimate the average treatment effect in a target population. (Westreich et al., 2017; Cole and Stuart, 2010). Thus, the assumption that the misclassification rates are the same in the training and test data can be viewed as a transportability assumption.

Implicit in the transportability assumption is that $p_{tr}(\mathbf{x}|y) = p_{test}(\mathbf{x}|y)$ (Pérez-Gállego et al., 2019), although the marginal distribution of the outcome in the training data, $p_{tr}(y)$, is allowed to be different from $p_{test}(y)$. This implies that quantification is best suited for $y \rightarrow \mathbf{x}$ problems (Fawcett and Flach, 2005) where the joint distribution of $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$. These type of problems occur when the latent outcome of interest, such as the presence or absence of a specific disease, causes distinct symptoms (Fawcett and Flach, 2005). Thus, even though we use $\mathbf{x}$ to predict y, $\mathbf{x}$ is only observed as a result of the causal chain beginning with y.

Under this transportability assumption,the conditional distribution of the predicted labels $\mathbf{a}$ from a classification algorithm is given by

$$p(\mathbf{a} \mid y) = \int_{\mathbf{x}} p(\mathbf{a} \mid \mathbf{x}) p(\mathbf{x} \mid y) d\mathbf{x} \,. \tag{4.1}$$

Here, $p(\mathbf{a} \mid \mathbf{x})$ is the prediction distribution from the algorithm, and is going to be same in the training and test sets for the same $\mathbf{x}$. Hence, if we assume that $p_{tr}(\mathbf{x}|y) = p_{test}(\mathbf{x}|y)$, then we have $p_{tr}(\mathbf{a} \mid y) = p_{test}(\mathbf{a} \mid y)$ in (4.1). That is, we assume that the sensitivity and specificity of the classifier is same in the training and test dataset.

Dataset shift occurs when both $p_{tr}(y) \neq p_{test}(y)$ and $p_{tr}(\mathbf{x}|y) \neq p_{test}(\mathbf{x}|y)$ (Moreno-Torres et al., 2012). It is evident from (4.1) that under dataset shift, we

will not generally have $p_{tr}(\mathbf{a} \mid y) = p_{test}(\mathbf{a} \mid y)$. This renders the assumptions of same sensitivity and specificity among the training and test sets used by most quantification methods invalid. An example of dataset shift is in the Population Health Metrics Research Consortium (PHMRC) gold standard dataset (Murray et al., 2011), which contains 168 reported symptoms and gold-standard underlying causes of death for adults in four countries. There are 21 total causes of death, that are then aggregated to 5 broader cause of death categories. Figure 4.1 shows the percentage of subjects within each country and cause of death that report each symptom. The x-axis is an enumeration of the entire list of symptoms x and the y-axis plots $p(x \mid y)$ for each symptom x. With no dataset shift, we would expect the conditional response rates (within each cause of death) for each question to be the same for each country. However, as the country-specific lines are quite distinct, it is clear that even within the same cause of death the reported symptom rate $p(\mathbf{x} \mid y)$ differs by country. This leads to poor performance when using symptoms and cause of death labels from 3 countries to predict the cause of death distribution for the remaining country (McCormick et al., 2016).

When limited data with known labels is available from the test set, Datta et al., 2018 have previously developed a quantification approach to address dataset shift called population-level Bayesian Transfer Learning (BTL) which resourcefully combines this limited labeled data with the predicted labels for all test data. The labeled test data, rather than the training data, are used to estimate the misclassification rates (sensitivities and specificities) of the

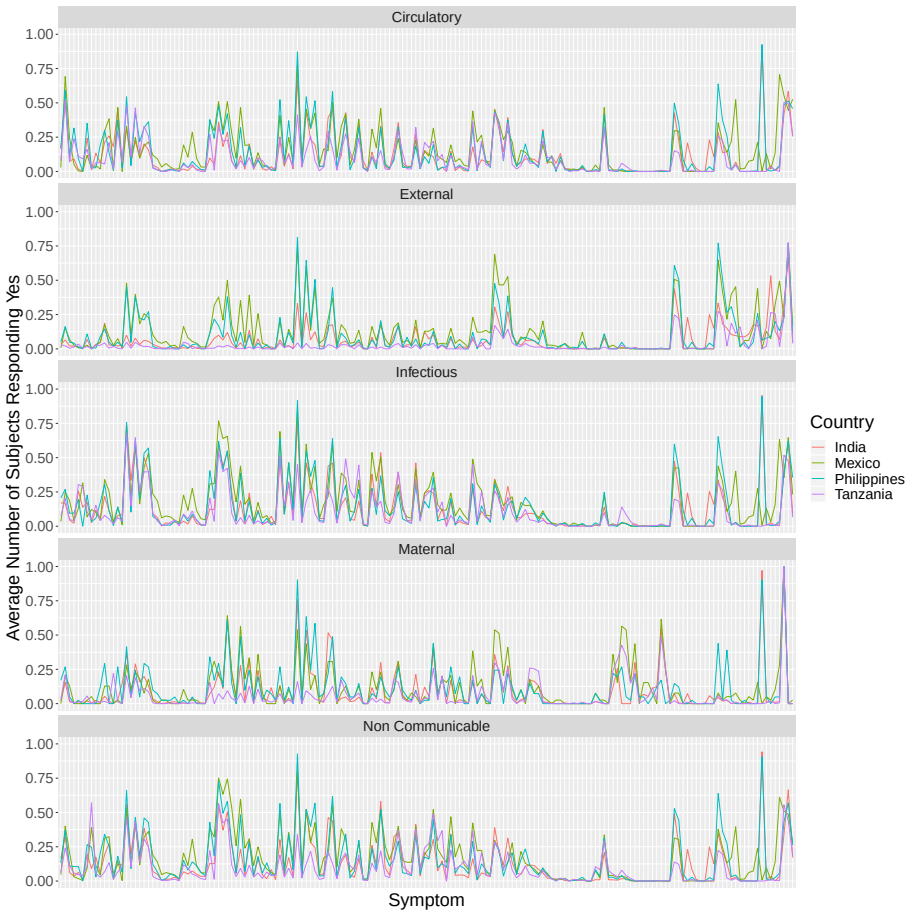


Figure 4.1: Percent of subjects with each of 168 reported symptoms within each of the 5 gold-standard underlying causes of death, by country.

algorithm on the test set. Importantly, only the distribution of $\mathbf{x}|y$ in the labeled instances is required to be representative of the whole test set. This is equivalent to assuming the misclassification rates are transportable from the labeled test data to the unlabeled test data. The marginal distribution of y in the labeled test set is allowed to be different from that in the unlabeled test set. Aside from its demonstrated performance in improving population level class estimates, the advantages of this approach are its simplicity, incorporating

multiple classification algorithms through an ensemble approach, and shrinkage to the default quantification estimates that assume perfect sensitivity and specificity when no labeled test data is available. In addition, this method only requires predictions from the available classifiers. It does not require access to the training data or additional training of prediction models using the labeled data, which can be a complicated endeavor given the high-dimensionality of the symptom space. However, quantification under dataset shift using BTL has three gaps that we aim to address here.

First, BTL requires a single-class prediction for each instance. Statistical classifiers are often probabilistic (McCullagh and Nelder, 1989; Murphy et al., 2006; Specht, 1990) producing the vector of prediction probabilities for every class. Hence, an additional step must be taken to convert this multi-class prediction distribution that is compositional data to a single-class prediction, a categorical data, by using some cutoff rule. Typically the most probable category is used, although cutoffs may have to be developed through cross-validation for algorithms such as Random Forest (Dahinden, 2011). Using this leads to information loss and Bella et al., 2010 showed that quantification using class probability estimates can outperform quantification using categorization. Thus, it is desirable to incorporate the entire compositional prediction distributions, instead of the single class predictions. Second, BTL does not allow for uncertainty in the true labeled test instances. Label uncertainty is not uncommon. For example, physicians may be uncertain in the final cause of death (McCormick et al., 2016), or labels may be produced by aggregating crowd sourced responses (Bragg, Weld, et al., 2013). Third, there

is no supporting theory about the accuracy of Bayesian quantification under dataset shift in large sample settings.

In this manuscript, we generalize Bayesian quantification using limited labeled test data to use entire prediction distributions from classifiers. To do so, rather than positing a valid likelihood for the compositional probability predictions, we derive Bayesian-style estimating equations derived from Kullback-Liebler divergence loss. The advantages of using this loss function over proper likelihoods for compositional data are many fold. The loss function is defined by a first moment (expectation) assumption and is robust to model misspecification. The loss function for the labeled data based on the conditional expectation is coherent with that for the unlabeled data based on the corresponding marginal expectation. Unlike Dirchlet higher-order distribution models for compositional data, the loss function approach allows 0's and 1's in the data. Also, importantly, this loss function remains the same no matter if one uses categorical single-class predictions or if one uses compositional probability predictions, subsuming the BTL model as a special case, when all data are categorical. The loss function harmonizes with conjugate priors for the parameters and a simple coarsening and rounding approximation leads to a fast and efficient Gibbs' sampler.

Next, we extend our approach to allow for probabilistic true labels. We use simple belief-based mixture modeling (Szczurek et al., 2010) to allow practitioners to specify the apriori class probabilities for instances in the labeled set.

Like BTL, our approach can combine multiple classifiers to produce an

ensemble quantification that is robust to inclusion of poor classifiers in the group. This also enables using two-versions of the output from the same classifier (the entire probability predictions or the most probable category). This is often helpful as there is some evidence suggesting that occasionally thresholding can perform better than using full predictive distributions (Quevedo, Luaces, and Bahamonde, 2012; Byass et al., 2012). Since it is not known apriori which choice of output will lead to more accurate quantification, the ensemble approach guards against the worst of the two choices for any dataset.

We demonstrate how different choices of shrinkage priors ensures that, in the absence of labeled test data, i.e., when it is not possible to adjust for dataset shift, quantification from our method shrinks to different existing quantification methods like classify & count (CC) (Forman, 2005) or probabalistic average (PA) (Bella et al., 2010). When using multiple classifiers, ensemble quantification from our approach, in absence of labeled test data, shrinks to the average of quantification over the set of classifiers using CC or PA.

Bayesian updating of posteriors using loss-functions is termed *generalized Gibbs updates* or *generalized belief updates*. The seminal work of Bissiri, Holmes, and Walker, 2016 explains the interpretation and statistical properties of such *generalized posteriors*. An immediate consequence of their work is that our loss-function, in an asymptotic sense, can be interpreted as a sum of two Bayes risks, one for the labeled data used to adjust for dataset shift and one for the unlabeled data to perform quantification. Going beyond this nice interpretation, we prove a theoretical guarantee about the asymptotic consistency of our quantification approach. The theory does not require full

specification of a true model and only relies on the first-moment assumption being true for some parameter value. The theory extends easily to the case of multiple classifiers.

Because our model handles both single-class and probabilistic predictions from a classifier, in addition to probabilistic true labels, and uses generalized Gibbs updates, we term it *Generalized Bayesian Quantification Learning (GBQL)*. Despite not positing a parametric likelihood for the compositional individual predictions, we develop and justify a simple and fast Gibbs sampler for obtaining posterior samples for the parameters of interest. We show the robustness of our method through simulations, and demonstrate its performance on the problem of deriving the cause-specific rates of child death using the PHMRC dataset.

4.2 Notation, assumptions, and review of quantification learning

We have N instances in our test set with predicted labels $\mathbf{a} = \mathbf{a}(\mathbf{x})$ output from a pre-trained algorithm A, but without the true labels y. The instances are assumed to be randomly sampled from our population of interest and our interest lies in estimating the distribution of y. We further assume availability of $n \ll N$ instances from our population of interest with both true labels y and predicted labels $\mathbf{a}$. We do not assume that the training data for the algorithm is available, nor do we assume the knowledge of the covariates $\mathbf{x}$ for the test set, as long as $\mathbf{a}(\mathbf{x})$ is available to us. Because true labels are potentially expensive to obtain, n is typically much smaller than N (and potentially n can be zero at

the beginning of a quantification project like burden of disease estimation in a country), so even if the covariates were available, the limited labeled data is not sufficient for building a new classifier as the covariate vectors $\mathbf{x}$ are typically high-dimensional.

We refer to the population from which we obtain unlabeled instances as $\mathcal{U}$ and the sub-population from which we obtain labeled instances as $\mathcal{L}$. Although $\mathcal{L}$ is a subset of the same test population, we do not require the distribution of y in $\mathcal{L}$ to be representative of our whole population. This is because true labels for outcomes may only be available for a convenient sample. For example, true cause of death may only be diagnosed for individuals who die in settings such as a hospital, making it impossible to also randomly sample individuals with known labels from our population of interest. We only assume that the conditional distribution $p(\mathbf{x} \mid y)$ is the same in the labeled and unlabeled instances. This transportability assumption for $p(\mathbf{x} \mid y)$ between $\mathcal{L}$ and $\mathcal{U}$ is more likely to hold. For example, even if the marginal cause of death distributions are different for hospital and community deaths, given a cause y, the symptoms $\mathbf{x}$ observed in the patient are likely to have similar distribution in both settings. The transportability assumption implies from (4.1) that $p(\mathbf{a} \mid y)$ is also transportable between $\mathcal{L}$ and $\mathcal{U}$ as the classifier $p(\cdot \mid \mathbf{x})$ is learnt from training data and this distribution remains same given $\mathbf{x}$ irrespective of the population $\mathbf{x}$ is drawn from.

We let $y_r \in \{1, \ldots, C\}$ denote the true class (label) for each instance r where C is the the total number of categories. Our target of interest is $\mathbf{p} = p_{\mathcal{U}}(y) = (p_1, \ldots, p_C)'$, the distribution of the outcome y in our population of interest

$\mathcal{U}$, i.e, $p_i = p(y_r = i | r \in \mathcal{U})$. An algorithm has been trained using labeled training data that produces a compositional score $\mathbf{a}(\mathbf{x}_r) = \mathbf{a}_r = (a_{r1}, \ldots a_{rC})$ for an instance r with covariate $\mathbf{x}_r$ such that $0 \leq \mathbf{a}_r \leq 1$ and $\sum_{i=1}^{C} a_{ri} = 1$. These scores may be an actual estimate of $p(y_r = i|\mathbf{x}_r)$, or simply a normalized degree of belief about whether $y_r = i|\mathbf{x}_r$. If a classifier gives a single predicted class j for an instance, in which case we would have $a_{rj} = 1$ and $a_{rj'} = 0$ for $j' \neq j$. Note that because these scores are produced via the training data, these are only expected to be accurate in the $r \in$ training data, and not for $r \in \mathcal{U} \cup \mathcal{L}$.

The most simple quantification approach is called Classify & Count (CC) (Forman, 2005). CC requires that there is a single predicted class j for each instance, so that $a_{rj} \in \{0, 1\}$. The CC estimate of p_i is simply

$$\hat{p}_i^{CC} = \frac{\sum_{r \in \mathcal{U}} a_{ri}}{N}.$$

An Adjusted Classify & Count (ACC) (Forman, 2005) method has been proposed to account for the fact that a classification algorithm is not expected to make perfect predictions, even for instances from the same population as the training data. ACC relies on cross-validation with the original training data to estimate the true positive and false positive rates (tpr and fpr) of the classifier (for the base case of $C = 2$), and obtaining the following ACC estimate of p_i

$$\hat{p}_i^{ACC} = \frac{\hat{p}_i^{CC} - fpr}{tpr - fpr}. \tag{4.2}$$

This method and its multi-class extensions (Hopkins and King, 2010) are inappropriate for quantification in the presence of dataset shift, as the fpr and tpr estimated from the training data will not be representative of the true fpr and tpr in the test population $\mathcal{U} \cup \mathcal{L}$ (Pérez-Gállego et al., 2019). Furthermore, $\hat{p}_i^{ACC}$ can be outside of the restricted range of $[0, 1]$, although Hopkins and King, 2010 correct for this using constrained optimization.

Bayesian Transfer Learning (BTL) (Datta et al., 2018) first proposes a model-based version of Classify and Count as $\sum_{r \in \mathcal{U}} \mathbf{a}_r \sim Multinomial(N, \mathbf{p}^{CC})$ and then adjusts for dataset shift. The adjustment follows from the simple observation that $\mathbf{p}^{CC}$ is actually $p_{\mathcal{U}}(\mathbf{a})$ and does not necessarily equal $\mathbf{p} = p_{\mathcal{U}}(y)$. In fact, the two are related by the identity $\mathbf{p}^{CC} = \mathbf{M}'\mathbf{p}$ where $\mathbf{M} = (M_{ij}) = (p(\mathbf{a}_r = j \mid y_r = i, r \in \mathcal{U} \cup \mathcal{L}))$ is the misclassification matrix of the classifier on the test population. This adjustment is conceptually the same as the one used by ACC and Hopkins and King, 2010. Instead of using $\mathbf{M} = \mathbf{I}$ (i.e., no adjustment as in CC) or $\mathbf{M} = \mathbf{M}_{tr}$ (i.e., transportability of the conditional distributions between the training and test data as used in ACC), BTL estimates $\mathbf{M}$ using data from $\mathcal{L}$, i.e., only assumes transportability of the conditional distributions from the limited test subset $\mathcal{L}$ to all test data. BTL does not assume any transportability of the marginal distribution of y between $\mathcal{L}$ and $\mathcal{U}$. The joint Bayesian hierarchical framework is then specified as

$$\begin{gathered} \sum_{r \in \mathcal{U}} \mathbf{a}_r \sim Multinomial(N, \mathbf{M}'\mathbf{p}) \\ \mathbf{a}_r \mid y_r = i \overset{ind}{\sim} Multinomial(1, \mathbf{M}_{i*}) \text{ for } r \in \mathcal{L}, i = 1, \ldots, C., \end{gathered} \tag{4.3}$$

with $\mathbf{M}_{i*}$ denoting the i^{th} row of $\mathbf{M}$. The advantages of the Bayesian frameworks are a) for any prior on $\mathbf{p}$ supported on the C-dimensional simplex (like

a Dirichlet distribution) the posterior is also guaranteed to lie on the simplex unlike ACC, and b) use of shrinkage priors for $\mathbf{M}$ to stabilize estimation when $\mathcal{L}$ is very small. The Bayesian setup also seamlessly allows for extensions like use of predictions from multiple classifiers, and allowing $\mathbf{M}$ and $\mathbf{p}$ to be a function of covariates.

Bella et al., 2010 developed approaches to quantification similar to CC and ACC, but using probabilistic classifiers, i.e., $\mathbf{a}_r$ being a compositional outcome instead of a categorical outcome. The Probabilistic Average (PA) estimate of p_i, $\hat{p}_i^{PA}$, is obtained in the same manner as $\hat{p}_i^{CC}$, but does not require $a_{rj} \in \{0, 1\}$. An adjusted version of the PA estimate (APA) uses probabilistic estimates of the tpr and fpr by taking the average estimated probability within each class; this is easily extended to 3 or more classes. However, like CC and ACC, they do not adjust for dataset shift. To our knowledge, there is no quantification method for dataset shift that utilizes the compositional predictions from probabilistic classifiers. Given the advantages of the BTL approach to quantification under dataset shift using limited labeled test data, we propose a generalization that can use both categorical and compositional predictions from classifiers.

4.3 Method

4.3.1 Issues with Bayesian quantification using compositional labels

There are fundamental hurdles to extend the model in (4.3) when some or all $\mathbf{a}_r$ are compositional. The Dirichlet distribution and its generalizations (Hijazi

and Jernigan, 2009; Wong, 1998; Tang and Chen, 2018), are the standard model for compositional data. However, there are several issues with specifying a Dirichlet model for $\mathbf{a}_r$.

1. We allow the $\mathbf{a}_r$ to take 0 and 1 values for the same dataset, as some $\mathbf{a}_r$ may be compositional while the remaining can be categorical. Classifiers that have an in-built thresholding rule for eliminating classes with small prediction probability will yield such mixed data types. Dirichlet distributions doesn't support 0's and 1's. and would require forcing the a_{rj}'s to lie strictly in $(0, 1)$ using a cutoff. The choice of such a cutoff is arbitrary. Alternatively, one can use the zero-inflated Dirichlet distribution (Tang and Chen, 2018) to formally account for the presence of 0's, which leads to a significant increase in the number of parameters.

2. The BTL approach (4.3) has a coherence property. The conditional model in the bottom-row $\mathbf{a}_r \mid y = i \sim Multinomial(1, \mathbf{M}_{i*})$ for $r \in \mathcal{U} \cup \mathcal{L}$ leads to the marginal model in the top row $\mathbf{a}_r \sim \sum_{i=1}^{C} p_i Multinomial(1, \mathbf{M}_{i*}) = Multinomial(1, \mathbf{M}'\mathbf{p})$ for $r \in \mathcal{U}$. Specifying $\mathbf{a}_r \mid y = i$ as a Dirichlet distribution (or its variants), will endow $\mathbf{a}_r$ with a mixture-Dirichlet marginal distribution which presents a computational challenge in posterior sampling.

3. Alternatively, one can enforce coherence in the conditional and marginal expectations by specifying models of the form $\mathbf{a}_r \mid y = i \sim Dirichlet(\alpha_1 \mathbf{M}_{i*})$ and $\mathbf{a}_r \sim Dirichlet(\alpha_2 \mathbf{M}'\mathbf{p})$. Such Dirichlet models for the data is susceptible to model misspecification. While more complex models like

generalized Dirichlet (Wong, 1998) can be used, increased model complexity usually comes with added computational burden. In addition, specifying a distribution would be specific to the classifier, training data, and test data, and would be very time-consuming and prone to error. Misspecification of the likelihood can lead to incorrect inference for $\mathbf{p}$, which can make the dataset-shift adjusted estimate of $\mathbf{p}$ even worse than the unadjusted one.

4. Single-class classifiers can be viewed as a subclass of probabilistic classifiers, with the predicted distribution being degenerate. Hence, if using two classifiers, one with compositional predictions and one single-class predictions, use of the Dirichlet model for the former and a multinomial model for the latter is discordant.

5. The multinomial likelihood for $\mathbf{a}_r$ nicely harmonizes with conjugate Dirichlet priors for the parameters $\mathbf{M}$ and $\mathbf{p}$ leading to an extremely efficient Gibbs sampler. Using a Dirichlet distribution based likelihood relinquishes this computational advantage as the priors no longer remain conjugate.

Finally, as an alternate to Dirichlet-based likelihoods, one can transform the data and use log-ratio models, which uses a multivariate normal or skew-normal to model the log-ratio coordinates of the compositional $\mathbf{a}_r$ (Comas-Cufí, Martín-Fernández, and Mateu-Figueras, 2016). However, a transformation-free approach is generally more desirable. Also, a model on the transformed compositional $\mathbf{a}_r$ will be discordant with the multinomial model for the categorical $\mathbf{a}_r$. The transformations also generally do not allow for 0's and 1's.

4.3.2 Bayesian estimating equations for compositional data

Central to BTL's estimation of population class probabilities ("quantification")

$$p(y_r = i) = p_i, \ \forall r \in \mathcal{U} \tag{4.4}$$

(4.3) is the assumption of transportability of conditional distribution between $\mathcal{L}$ and $\mathcal{U}$, i.e.,

$$p(\mathbf{a}_r \mid y_r = i) = \mathbf{M}_{i*} \ \forall r \in \mathcal{U} \cup \mathcal{L}. \tag{4.5}$$

The distributional assumption (4.5) can also be viewed as a first-moment assumption

$$E(\mathbf{a}_r \mid y_r = i) = \mathbf{M}_{i*} \ \forall r \in \mathcal{U} \cup \mathcal{L}. \tag{4.6}$$

The two viewpoints are equivalent for categorical $\mathbf{a}_r$ used in BTL, but (4.6) is more general as it is no longer restricted to categorical data. For compositional $\mathbf{a}_r$, rather than specifying $p(\mathbf{a}_r|y_r = i)$, we only make the general first moment assumption (4.6). This is similar to the first-moment assumption in the PA and APA approaches. The challenge is of course how to do valid Bayesian inference without a full model specification.

First focusing on labeled instances $r \in \mathcal{L}$, we consider the following loss function to connect the parameter $\mathbf{M}$ to our data $\mathbf{a}_r, y$

$$\ell_{\mathcal{L}}(\mathbf{M} \mid \{\mathbf{a}_r, y_r\}_{r\in\mathcal{L}}) = \sum_{r\in\mathcal{L}} D_{KL}(\mathbf{a}_r || \sum_{i=1}^{C} \mathbf{M}_{i*} I(y_r = i)) \tag{4.7}$$

where $D_{KL}(\mathbf{p}||\mathbf{q})$ is the Kullback–Leibler divergence (KLD) between two

distributions $\mathbf{p}$ and $\mathbf{q}$. There are several reasons to choose the KLD loss functions. First, if (4.6) is true for some $\mathbf{M} = \mathbf{M}_0$, then

$$E_{\mathbf{M}_0}\left(\frac{d\ell_{\mathcal{L}}}{d\mathbf{M}}\right) = 0\ . \tag{4.8}$$

To see this, observe that $d\ell_{\mathcal{L}}/d\mathbf{M}$ is the derivative of a multinomial likelihood. Hence, $E_{\mathbf{M}_0}(d\ell_{\mathcal{L}}/d\mathbf{M}) = 0$ when $\mathbf{a}_r$ are categorical. However, this derivative is only a linear function of $\mathbf{a}_r$ and hence the expectation remains unchanged when we switch to compositional $\mathbf{a}_r$ with the same conditional mean. Hence, the loss function ℓ_L leads to a set of unbiased estimating equations (Liang and Zeger, 1986) for compositional data. The second advantage of using KLD is that, as $x \log x = 0$, it seamlessly accommodates instances 0's and 1's in $\mathbf{a}_r$. Most importantly, minimizing (4.7) is equivalent to maximizing

$$\prod_{r\in\mathcal{L}}\prod_{j=1}^{C}\left(\sum_{i} I(y_r = i)M_{ij}\right)^{a_{rj}}$$

which is the exact form of the multinomial quasi-likelihood (MQL). So, when $\mathbf{a}_r$ are all categorical, this reduces to the likelihood from the second row of (4.3).

If only inference on $\mathbf{M}$ was of interest, frequentist optimization on (4.7) or GEE using its derivative can be executed. Using the rich theory of estimating equations, the estimate $\widehat{\mathbf{M}}$ has been shown to be a consistent estimator for $\mathbf{M}$ (Papke and Wooldridge, 1996; Mullahy, 2015), and such frequentist approaches have been commonly used in the econometrics literature for regression with a compositional outcome.

However, the primary interest in quantification is in $\mathbf{p}$ and accurate estimation of the nuisance parameter $\mathbf{M}$ is only an important intermediate step. The unlabeled dataset $\mathcal{U}$ is the only one informing estimation of $\mathbf{p}$, and using (4.4) and (4.6), the marginal first-moment condition for $\mathbf{a}_r$ in $\mathcal{U}$ is given by:

$$E[\mathbf{a}_r] = E[E[\mathbf{a}_r|y_r]] = \sum_i p_i E[\mathbf{a}_r|y_r = i] = \mathbf{M}'\mathbf{p}, \forall r \in \mathcal{U}. \tag{4.9}$$

This harmonizes with the loss-function

$$\ell_{\mathcal{U}}(\mathbf{p}, \mathbf{M} \mid \{\mathbf{a}_r\}_{r\in\mathcal{U}}) = \sum_{r\in\mathcal{L}} D_{KL}(\mathbf{a}_r||\mathbf{M}'\mathbf{p}) \ . \tag{4.10}$$

The loss function $\ell_{\mathcal{U}}$ for the marginal distribution of the predicted labels is coherent with the loss-function $\ell_{\mathcal{L}}$ for their conditional distribution, as they are based off of coherent moment conditions (4.6) and (4.9). Assuming (4.4) and (4.6) holds for some true $\mathbf{p}_0$ and $\mathbf{M}_0$, following the same logic used in (4.8), we can show

$$E_{\mathbf{M}_0,\mathbf{p}_0}\left(\frac{d\ell_{\mathcal{U}}}{d(\mathbf{M},\mathbf{p})}\right) = 0, \tag{4.11}$$

i.e., the derivative is once again an estimating equation. However, if we only considered $\ell_{\mathcal{U}}$ without bringing in $\ell_{\mathcal{L}}$, $\mathbf{M}$ and $\mathbf{p}$ cannot be identified. For example, $\ell_{\mathcal{U}}(\mathbf{M}, \mathbf{p}) = \ell_{\mathcal{U}}(\mathbf{I}, \mathbf{M}'\mathbf{p})$. Hence, we will consider the joint loss-function $\ell_{\mathcal{L}} + \ell_{\mathcal{U}}$ as adding $\ell_{\mathcal{L}}$ helps to identify $\mathbf{M}$ which in turns makes $\mathbf{p}$ identifiable.

Loss functions and estimating equations have traditionally been used in frequentist literature and have been shown to yield inference robust to model misspecification. To conduct and justify Bayesian inference with loss functions, we invoke the fundamental results of Bissiri, Holmes, and Walker, 2016 who

showed that for any reasonable choice of a loss-function $\ell(\boldsymbol{\theta} \mid data)$ and prior $\Pi(\boldsymbol{\theta})$, generalized Gibbs posteriors of the form

$$\Pi(\boldsymbol{\theta} \mid data) \propto \exp\left(-\ell(\boldsymbol{\theta} \mid data)\right) \Pi(\boldsymbol{\theta})$$

are valid posteriors provided the normalizing constant exists. This posterior is interpreted as the distribution ν for $\boldsymbol{\theta}$ minimizing the loss function $E_\nu(\ell(\boldsymbol{\theta} \mid data)) + D_{KL}(\nu, \Pi)$.

We will use the notation $\mathbf{a}^{\mathcal{L}}$ and $\mathbf{a}^{\mathcal{U}}$ to respectively denote the collections $\{\mathbf{a}_r\}_{r\in\mathcal{L}}$ and $\{\mathbf{a}_r\}_{r\in\mathcal{U}}$, and similarly for collections of the other variables. The two-loss functions ℓ_L and ℓ_U also have same functional form leading to the generalized posterior:

$$\begin{aligned}\Pi(\mathbf{p}, \mathbf{M} \mid \mathbf{a}^{\mathcal{U}}, \mathbf{a}^{\mathcal{L}}, y^{\mathcal{L}}) &\propto \exp\left(-\sum_{r\in\mathcal{U}} D_{KL}(\mathbf{a}_r || E[\mathbf{a}_r]) - \sum_{r\in\mathcal{L}} D_{KL}(\mathbf{a}_r || E[\mathbf{a}_r|y_r])\right) \Pi(\mathbf{p}, \mathbf{M}) \\ &\propto \exp\left(\sum_{r\in\mathcal{U}}\sum_{j=1}^{C} a_{rj} \log \frac{\sum_i p_i M_{ij}}{a_{rj}} + \sum_{r\in\mathcal{L}}\sum_{j=1}^{C} a_{rj} \log \frac{\sum_{i=1}^{C} I(y_r = i) M_{ij}}{a_{rj}}\right) \Pi(\mathbf{p}, \mathbf{M})\end{aligned}$$

If all $\mathbf{a}_r$ were categorical, this posterior is identical to the one from the BTL model (4.3). However, using the estimating equations approach, we now have an unified framework for Bayesian quantification for both categorical, compositional or mixed-type $\mathbf{a}_r$ without having to specify the full models for the different data types.

4.3.3 Uncertainty in true labels

As stated in Section 4.1, in many applications, there is uncertainty in some or all of the true labels in the labeled test set $\mathcal{L}$. For example, a panel of physicians

may fail to unanimously agree on a single cause, and may provide a subset of the list of causes from which they believe the individual was equally likely to die. In this Section, we modify the loss function $\ell_{\mathcal{L}}$ to incorporate uncertainty for class labels in $\mathcal{L}$.

Following the belief based modeling framework of Szczurek et al., 2010, we let b_{ri} represent the apriori probability that instance r belongs to label i. $\mathbf{b}_r$ is constrained such that $0 \leq b_{ri} \leq 1$ and $\sum_{i=1}^{C} b_{ri} = 1$. Now for an instance $r \in \mathcal{L}$ we no longer observe the y_r's but observe the belief vector $\mathbf{b}_r$. Cases where the true label is identified with complete certainty can be subsumed by writing $\mathbf{b}_r = \mathbf{e}_i$ when $y_r = i$, $\mathbf{e}_i$ denoting the vector with 1 at the i^{th} component and zeros elsewhere. We can generalize the conditional first-moment condition (4.6) to

$$E[\mathbf{a}_r|\mathbf{b}_r] = E[E[\mathbf{a}_r|y_r, \mathbf{b}_r] \mid \mathbf{b}_r] = E\left(\sum_i M_{i*} I(y_r = i) \mid \mathbf{b}_r\right) = \mathbf{M}'\mathbf{b}_r$$

and our loss function for $\mathcal{L}$ becomes

$$\ell_{\mathcal{L}}(\mathbf{M} \mid \{\mathbf{a}_r, \mathbf{b}_r\}_{r\in\mathcal{L}}) = -\sum_{r\in\mathcal{L}} D_{KL}(\mathbf{a}_r||\mathbf{M}'\mathbf{b}_r) = \sum_{r\in\mathcal{L}}\sum_{j=1}^{C} a_{rj} \log\left(\frac{\sum_{i=1}^{C} b_{ri} M_{ij}}{a_{rj}}\right) \tag{4.12}$$

Of course, the loss for the unlabeled data remains the same, and Bayesian inference proceeds using the likelihood $\ell_{\mathcal{L}} + \ell_{\mathcal{U}}$ with this generalized choice of $\ell_{\mathcal{L}}$.

Once again, appealing to the results from Bissiri, Holmes, and Walker, 2016, we can see that $\nu = \Pi(\mathbf{p}, \mathbf{M}|\mathbf{a}^{\mathcal{U}}, \mathbf{a}^{\mathcal{L}}, \mathbf{b}^{\mathcal{L}})$ is the probability measure which, as

$n, N \to \infty$ and $\frac{n}{N} \to \alpha$, minimizes the Bayes risk

$$E_\nu \left[E_{r \in \mathcal{U}}[D_{KL}(\mathbf{a}_r || \mathbf{M}^{'} \mathbf{p})] + \alpha E_{r \in \mathcal{L}}[D_{KL}(\mathbf{a}_r || \mathbf{M}^{'} \mathbf{b}_r)] \right] .$$

4.3.4 Ensemble Quantification Incorporating Multiple Predictions

There may be $k = 1, \ldots, K$ predictions for each instance corresponding to predictions from different classifiers, such as logistic regression versus a support vector machine. Datta et al., 2018 has already shown the advantage of incorporating multiple algorithms for quantification when only categorical predictions are available, and their ensemble quantification can easily be extended to compositional settings.

We represent the k^{th} algorithm prediction for instance r as $\mathbf{a}_r^k$. A fundamental observation for the ensemble approach is that each algorithm is expected to have their own sensitivities and specificities. Hence, the conditional first moment assumption (4.6) becomes

$$E(\mathbf{a}_r^k \mid y_r = i) = \mathbf{M}_{i*}^k \ \forall r \in \mathcal{U} \cup \mathcal{L}. \tag{4.13}$$

For the unlabeled data, we will now have the labels satisfying the marginal first moment condition $E(\mathbf{a}_r^k = \mathbf{M}^{k'} \mathbf{p})$. Hence, each of the K predictions for the unlabeled test data $\mathcal{U}$ informs about the same parameter $\mathbf{p}$ and we can conduct ensemble quantification by specifying a loss function which is the sum of the losses for the individual algorithms:

$$\sum_{k=1}^{K}\left[\sum_{r\in\mathcal{U}} D_{KL}(\mathbf{a}_r^k||\mathbf{M}^{(k)'}\mathbf{p}) + \sum_{r\in\mathcal{L}} D_{KL}(\mathbf{a}_r^k||\mathbf{M}^{(k)'}\mathbf{b}_r)\right]$$

An advantage of this loss function is that it allows for combining information from probabilistic classifiers and non-probabilistic ones (like clinical classifiers for cause of deaths). Additionally, we can now also use multiple predictions from the same classifier but using a different output format, e.g., one using the full composition prediction distribution versus one only retaining the rescaled scores for the top-S classes and thresholding the rest to zero ($S = 1$ is the categorical prediction).

4.3.5 Gibbs Sampler using rounding and coarsening

We first outline the Gibbs sampler steps when only one predicted labels is available per instance. The sampler for ensemble quantification is detailed in the appendix. The generalized posterior distribution ν is given by

$$\nu \propto \left[\prod_{r\in\mathcal{U}}\prod_{j=1}^{C}\left(\sum_i p_i M_{ij}\right)^{a_{rj}} \prod_{r\in\mathcal{L}}\prod_{j=1}^{C}\left(\sum_i b_{ri} M_{ij}\right)^{a_{rj}}\right]\pi(\mathbf{p},\mathbf{M})$$

When all $\mathbf{a}_r$ are categorical, the polynomial expansion of $(\sum_i p_i M_{ij})^{\sum_r a_{rj}}$ enabled an efficient latent variable Gibbs sampler in Datta et al., 2018. When a_{rj} are fractions, this convenience is lost as fractional polynomials do not have such neat expansions. Additionally, since we now allow uncertainty in the true labels, we also need to consider the extra fractional expansion terms $(\sum_i b_{ri} M_{ij})^{a_{rj}}$.

To enable fast and efficient sampling, we first switch from ν to ν_{round} where

the probabilistic output a_{rj} is replaced by $\lceil Ta_{rj} \rceil$ where T is an integer, and $\lceil \cdot \rceil$ denotes the ceiling of any real number. Now, consider now the following generative model:

$$z_{rt} \overset{\text{ind}}{\sim} \begin{cases} Multinomial(1, \mathbf{p}) & \text{if } r \in \mathcal{U} \\ Multinomial(1, \mathbf{b}_r) & \text{if } r \in \mathcal{L} \end{cases}, \; t = 1, \ldots, T_r = \sum_j \lceil Ta_{rj} \rceil$$

$$d_{rt} | z_{rt} = i \overset{\text{ind}}{\sim} Multinomial(1, \mathbf{M}_{i*}), r \in \mathcal{L} \cup \mathcal{U}$$

The rounded generalized posterior ν_{round} is then the proper Bayesian posterior using the likelihood $p(\mathbf{d}^{\mathcal{U}}, \mathbf{d}^{\mathcal{L}} \mid \mathbf{b}^{\mathcal{L}}, \mathbf{M}, \mathbf{p})$ for any realization of $\mathbf{d}_{rt}$'s satisfying $\sum_t I(\mathbf{d}_{rt} = j) = \lceil Ta_{rj} \rceil$. To obtain samples of $\mathbf{p}$ and $\mathbf{M}$ from ν_{round}, instead of using this marginalized likelihood, we can equivalently introduce $\mathbf{z}^{\mathcal{L}}$, and $\mathbf{z}^{\mathcal{U}}$ as latent variables and use the joint likelihood $p(\mathbf{d}^{\mathcal{U}}, \mathbf{d}^{\mathcal{L}}, \mathbf{z}^{\mathcal{L}}, \mathbf{z}^{\mathcal{U}} \mid \mathbf{b}^{\mathcal{L}}, \mathbf{M}, \mathbf{p})$. This joint likelihood decomposes nicely and will be conducive to a Gibbs sampler with standard Dirichlet priors on $\mathbf{M}$ and $\mathbf{p}$.

Next, since we artificially inflate sample size by an order of T by switching from $\mathbf{a}_r$ to $\lceil T\mathbf{a}_r \rceil$, instead of sampling from ν_{round} we sample from the coarsened likelihood

$$\nu_{coarse} \propto p(\mathbf{d}^{\mathcal{U}}, \mathbf{d}^{\mathcal{L}} | \mathbf{b}^{\mathcal{L}}, \mathbf{M}, \mathbf{p})^{\frac{1}{T}} \pi(\mathbf{p}, \mathbf{M})$$

Because $p(\mathbf{d}^{\mathcal{U}}, \mathbf{d}^{\mathcal{L}} | \mathbf{b}^{\mathcal{L}}, \mathbf{M}, \mathbf{p})$ is a proper likelihood, this implies that ν can be expressed as a power posterior (Bhattacharya, Pati, Yang, et al., 2019; Ibrahim et al., 2015; Miller and Dunson, 2019), and as $p(\mathbf{d}^{\mathcal{U}}, \mathbf{d}^{\mathcal{L}} | \mathbf{b}^{\mathcal{L}}, \mathbf{M}, \mathbf{p})$ is a mixture of categorical distributions, we can introduce latent variables into our

Gibbs sampler by using the Conditional Coarsening Algorithm (Miller and Dunson, 2019) just like we would do for ν_{round}.

For outlining the Gibbs sampler steps, we use generic Dirichlet priors $\mathbf{M} \sim Dirichlet(\mathbf{V})$, i.e, $\mathbf{M}_{i*} \overset{ind}{\sim} Dirchlet(\mathbf{V}_{i*})$, and $\mathbf{p} \sim Dirichlet(\mathbf{v})$ where $\mathbf{V}$ and $\mathbf{v}$ respectively are matrix and vector of positive hyperparameters. Specific choices with desirable shrinkage properties are discussed in Section 4.3.6. This gives the following Gibbs updates:

$$\mathbf{z}_r|\cdot \sim \begin{cases} Multinomial\left(1, \frac{1}{\sum_i M_{ij} p_i}(M_{1j}p_1, \ldots, M_{Cj}p_C)\right), \; r \in \mathcal{U}, \; d_{rt} = j \\ Multinomial\left(1, \frac{1}{\sum_i M_{ij} b_{ri}}(M_{1j}b_{r1}, \ldots, M_{Cj}b_{rC})\right), \; r \in \mathcal{L}, \; d_{rt} = j \end{cases}$$

$$M_i|\cdot \sim Dir\left(\tilde{\mathrm{V}}_{i1}, \ldots, \tilde{\mathrm{V}}_{iJ}\right), \; \tilde{\mathrm{V}}_{ij} = \mathrm{V}_{ij} + \frac{1}{T}\left(\sum_{r \in \mathcal{U}, \mathcal{L}} \sum_{t=1}^{T} (I(d_{rt} = j) I(z_{rt} = i)\right)$$

$$p|\cdot \sim Dir\left(\tilde{\mathrm{v}}_1, \ldots, \tilde{\mathrm{v}}_C\right), \; \tilde{\mathrm{v}}_i = \mathrm{v}_i + \frac{1}{T} \cdot \left(\sum_{r \in \mathcal{U}} \sum_{t=1}^{T} I(z_{rt} = i)\right)$$

If there are hyper-parameters γ in $\mathbf{V}$ and $\mathbf{v}$ that need to be assigned a prior, they can be sampled using a Metropolis-Hastings step. We note that the full conditional distributions for the $z_{rt}|d_{rt} = j$ for $r \in \mathcal{U}$ are identical, which enables them to be jointly sampled. Furthermore, the z_{rt} for $r \in \mathcal{L}$ do not need to be updated if there is a i such that $b_{ri} = 1$. We find that setting $T = 100$ works well in practice, and that there is little information to be gained by finer coarsening.

4.3.6 Shrinkage towards default quantification methods

We now discuss how existing quantification approaches are special cases of GBQL with specific choices of degenerate priors for $\mathbf{M}$, and how we leverage this knowledge to construct shrinkage priors in data-scarce settings.

Quantification projects like burden of disease estimation using nation-wide surveys are often multi-year endeavors, and at the initial stages of such projects, $\mathcal{L}$, consisting of hospital deaths with clinically diagnosed causes, can be very small. With very limited labeled data, estimating both $\mathbf{M}$ and $\mathbf{p}$ precisely with vague priors is ill-advised as $\mathbf{M}$ involves $C(C-1)$ parameters. Hence, it is important to carefully choose priors that stabilize estimation of $\mathbf{M}$.

We first make the following observations for the scenario where $n = 0$, i.e., when there is no labeled test set to estimate dataset shift. Consider a sequence $\{\Pi_u(\mathbf{M}) \mid u = 1, 2, \ldots\}$ of priors for $\mathbf{M}$ such that Π_u converges in distribution to the degenerate prior at some pre-fixed transition matrix $\mathbf{M}^{pr}$. Then the posterior ν_u using the prior $\Pi(\mathbf{p})\Pi_u(\mathbf{M})$ converges in distribution to

$$\lim_{u\to\infty} \nu_u(\mathbf{p}) \propto \exp\left(-\sum_{r\in\mathcal{U}} D_{KL}(\mathbf{a}_r||\mathbf{M}^{pr\prime}\mathbf{p})\right)\Pi(\mathbf{p})\ .$$

If $\mathbf{M}^{pr} = \mathbf{I}$, then for any prior choice of $\mathbf{p}$, $\lim_{u\to\infty} \nu_u(\mathbf{p}) \propto Dirichlet(\sum_{r\in\mathcal{U}} \mathbf{a}_r)\Pi(\mathbf{p})$. In particular, if $\Pi(\mathbf{p}) = Dirchlet(\mathbf{0})$ or as $N \to \infty$, then $\lim_{u\to\infty} \nu_u(\mathbf{p}) = Dirichlet(\sum_{r\in\mathcal{U}} \mathbf{a}_r)$. For categorical $\mathbf{a}_r$, this result was proved in Datta et al., 2018, and shows that $E_{\lim_u \nu_u}(\mathbf{p}) = \mathbf{p}^{CC}$, i.e., using priors $\Pi_u(\mathbf{M})$ shrinking towards the degenerate prior at $\mathbf{I}$, inference from GBQL becomes identical to inference from Classify and Count (Forman, 2005) when there is no labeled

dataset. Analogously, for the same settings, when $\mathbf{a}_r$ are compositional, posterior mean from GBQL becomes identical to Probabilistic Average. Extending, the argument to the settings with multiple predictions, it is straightforward to see that $E_{\lim_u \nu_u}(\mathbf{p}) = 1/K \sum_{k=1}^{K} \mathbf{p}^{k,PA}$, i.e., the posterior mean from our ensemble classifier coincides with the average of the PA estimates for the K classifiers.

Alternatively, if the misclassification matrix $\mathbf{M}^{tr}$ for the training data is available and can be trusted for test data, one can use $\mathbf{M}^{pr} = \mathbf{M}^{tr}$. Then the posterior $\lim_{u\to\infty} \nu_u(\mathbf{p})$ coincides with the implicit likelihood in Adjusted Classify and Count (for categorical $\mathbf{a}_r$) and in Adjusted Probabilistic Average (when $\mathbf{a}_r$ are compositional). In fact, using $\Pi(\mathbf{M}) \approx \delta(\mathbf{M} = \mathbf{M}^{tr})$ in GBQL is a better implementation of ACC or APA, as the proper posteriors ensure that the estimate (posterior mode or mean) of $\mathbf{p}$ is guaranteed to be a vector of probabilities lying in $[0, 1]$. This is not assured in their current implementation based on a direct correction (4.2).

Hence, in absence of local labeled set, a prior for $\mathbf{M}$ concentrated around $\mathbf{I}$ or $\mathbf{M}^{tr}$, makes estimates from GBQL nearly coincide with these existing methods (Figure 4.2). Such classes of shrinkage priors for $\mathbf{M}$ are easy to construct. For example, the priors $\mathbf{M}_{i*} \sim Dirichlet(\gamma_{ui}(\mathbf{M}^{pr}_{i*} + \epsilon_u \mathbf{1}))$ concentrates around $\delta(\mathbf{M} = \mathbf{M}^{pr})$ if either $\epsilon_u \to 0$ or $\gamma_{ui} \to \infty$. When we will have small amounts of labeled data, using these shrinkage priors will make a bias-variance tradeoff yielding estimates with higher precision. The benefits of such shrinkage priors over non-informative priors have been demonstrated in Datta et al., 2018 in such settings. Finally as more and more labeled data is collected, in the

next section we show that any reasonable choice of prior (including all these shrinkage priors) leads to desirable asymptotic concentration of the posterior.

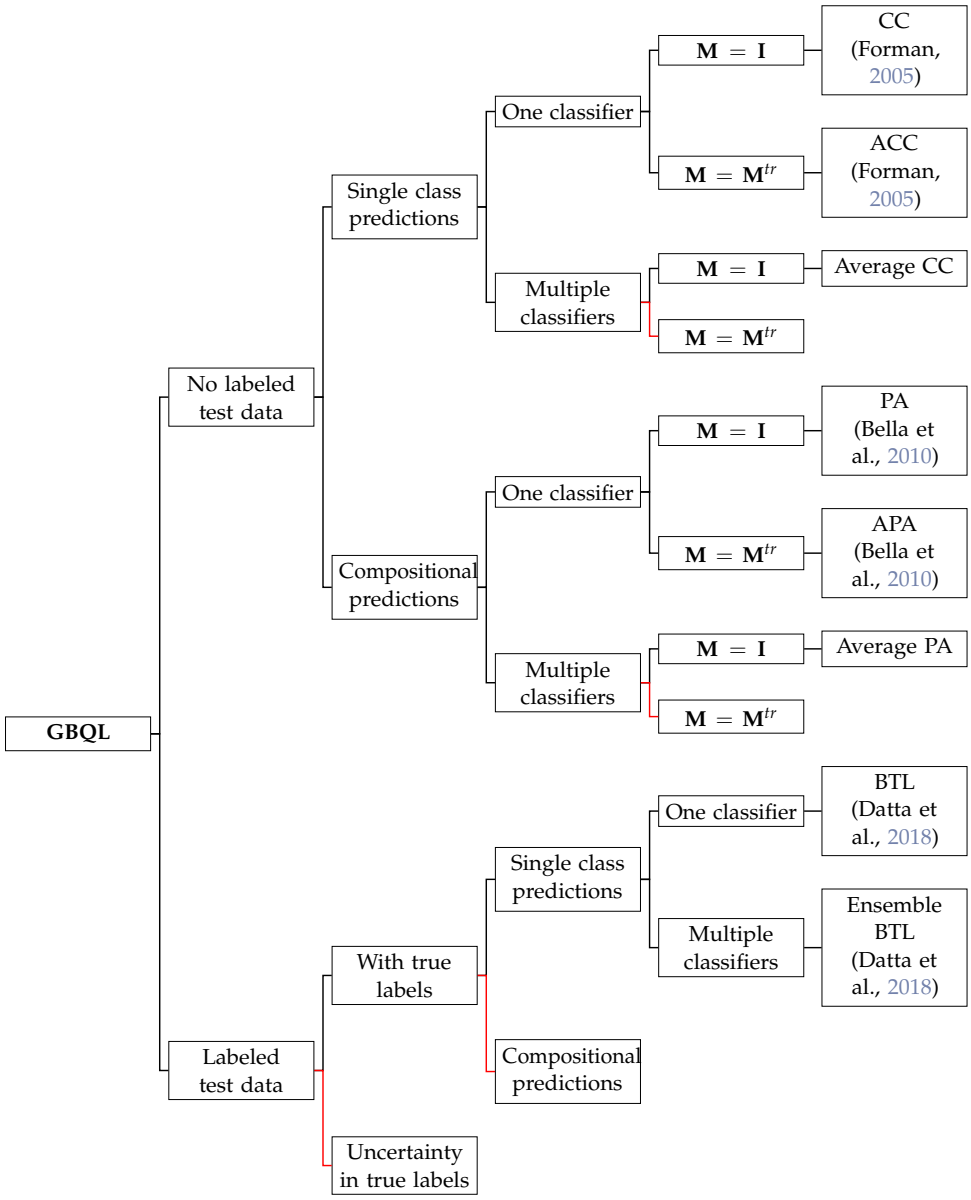


Figure 4.2: GBQL includes and extends the common quantification methods through different classifier outputs and choices of priors for $\mathbf{M}$. Red lines indicate where GBQL extends current methods, while black lines indicate where GBQL subsumes existing methods.

4.4 Theory

Our quantification approach is grounded in the only assumption that for both $\mathcal{L}$ and $\mathcal{U}$, the conditional first moment of $\mathbf{a_r} \mid y_r$ are correctly specified as in (4.6). Throughout we do not we do not make any other assumptions about higher moments or full distributions. Kessler and Munkin, 2015 have used a similar first moment assumption to develop a Gibbs sampling approach for compositional regression. However, their approach only incorporates the rounded likelihood for the psuedo-data $\mathbf{d}_r$ and does not coarsen. Rounding inflates the sample size by a factor of T resulting in underestimation of the posterior variance and the coarsening is needed to adjust for this. We will show that the coarsening adjustment by this factor of T ensures asymptotic equivalence of the rounded and coarsened posterior ν_{coarse} with the original posterior ν. Yuan, Chappell, and Bailey, 2007 also used a similar Gibbs sampler in the context of early-phase clinical trials with multiple toxicity grades. However, both Yuan, Chappell, and Bailey, 2007 and Kessler and Munkin, 2015 did not provide any theory backing the use of a Gibbs sampler based on a loss-function instead of a proper likelihood or justified the approximations used in the Gibbs sampler. Not only have we justified using the first-moment assumption in a Bayesian framework, by appealing to the results of Bissiri, Holmes, and Walker, 2016, in this section we also establish posterior consistency of both ν and the rounded and coarsened posterior ν_{coarse} used in the Gibbs sampler.

We develop the theory for the general case where the true labels in $\mathcal{L}$ are observed with uncertainty $\mathbf{b}_r$ which subsumes the case with exact labels y_r.

We will use $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{p}}$ to denote the free parameters in $\mathbf{M}$ and $\mathbf{p}$ respectively, i.e., $\widetilde{\mathbf{M}}$ excludes the last column of $\mathbf{M}$, $\widetilde{\mathbf{p}}$ excludes the last element of $\mathbf{p}$. $\mathbf{M}$ and $\mathbf{p}$ are bijective functions of $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{p}}$ respectively, so we will use them interchangeably. Let $\boldsymbol{\theta} = (\widetilde{\mathbf{M}}, \widetilde{\mathbf{p}})$, then $\boldsymbol{\theta}$ is supported on the compact set $\Theta = \mathcal{S}_{C-1}^{C} \otimes \mathcal{S}_{C-1}$ where $\mathcal{S}_d = \{\mathbf{x} \in \mathbb{R}^d \mid x_i \geq 0, \mathbf{1}'\mathbf{x} \leq 1\}$. Switching to $\widetilde{\mathbf{M}}$ and $\widetilde{\mathbf{p}}$ ensures that the parameter space Θ has a non-empty interior. The generalized posterior from Section 4.3.3 is given by:

$$\nu_N(\boldsymbol{\theta}) = \Pi(\mathbf{p}, \mathbf{M} \mid \mathbf{a}^{\mathcal{U}}, \mathbf{a}^{\mathcal{L}}, \mathbf{b}^{\mathcal{L}}) \propto \exp\left(- \sum_{r \in \mathcal{U}} D_{KL}(\mathbf{a}_r || \mathbf{M}'\mathbf{p}) - \sum_{r \in \mathcal{L}} D_{KL}(\mathbf{a}_r || \mathbf{M}'\mathbf{b}_r) \right) \Pi(\mathbf{p}, \mathbf{M}). \tag{4.14}$$

Let $\mathbf{p}^0$ and $\mathbf{M}^0$ denotes the true values and $\boldsymbol{\theta}^0 = (\widetilde{\mathbf{M}}^0, \widetilde{\mathbf{p}}^0)$, an interior point in Θ. We first state our assumptions, for the theory:

1. Let $\mathbf{B}_j$ denote the matrix of $\mathbf{b}_r$'s stacked as columns for $r \in \mathcal{L}$ such that $a_{rj} > 0$. Then $\mathbf{B}_j$ is full rank.

2. $\mathbf{M}^0$ is non-singular.

Theorem 4. *Let $B_\epsilon(\boldsymbol{\theta}^0)$ be the Euclidean ball of radius ϵ around $\boldsymbol{\theta}^0$, and $\Pi(\mathbf{p}, \mathbf{M})$ be any prior which gives positive support to $B_\epsilon(\boldsymbol{\theta}^0)$ for any $\epsilon > 0$. Then, under assumptions 1-2, as $N, n \to \infty$ and n/N to some limit, for any $\epsilon > 0$, $P_{\nu_N}(B_\epsilon(\boldsymbol{\theta}^0)) \to 1$.*

While the formal proof is provided in the appendix, we briefly outline the ideas used here which will also help to contextualize the assumptions. We can write $\nu_N(\boldsymbol{\theta}) \propto \exp(-\ell_{\mathcal{L},n}(\widetilde{\mathbf{M}}) - \ell_{\mathcal{U},N}(\boldsymbol{\theta}))\Pi(\widetilde{\mathbf{p}}, \widetilde{\mathbf{M}})$ where the subscript N is added to indicate dependence of $\ell_{\mathcal{L}}$, $\ell_{\mathcal{U}}$ and ν on the sample size. Recently, Miller, 2019 has provided very general and useful conditions for establishing asymptotic concentrations of generalized posteriors of the form

$\exp(-Nf_N(\boldsymbol{\theta}))\Pi(\boldsymbol{\theta})$. One of the general tricks is to show that the functions f_N converge point-wise to some function f, and that f_N's and f are convex. These conditions are sufficient for the generalized posterior to concentrate around $\boldsymbol{\theta}^0$, the minimizer of f. In our case, $f_N = (\ell_{\mathcal{L},n} + \ell_{\mathcal{U},N})/N$ converges point-wise to $f = \alpha E_{\mathcal{L}}(D_{KL}(\mathbf{a}||\mathbf{M}'\mathbf{b})) + E_{\mathcal{U}}(D_{KL}(\mathbf{a}||\mathbf{M}'\mathbf{p}))$ where $\alpha = \lim n/N$. However, neither f_N's nor f is convex because of the $\mathbf{M}'\mathbf{p}$ term, ruling out direct application of this result.

We hence first focus just on $\ell_{\mathcal{L},n}/n$ and establish the result

Lemma 1. *If assumption 1 holds, then for any $\epsilon > 0$ the generalized posterior $\nu_{\mathcal{L},n}(\widetilde{\mathbf{M}}) \propto \exp(-\ell_{\mathcal{L},n})\Pi(\widetilde{\mathbf{M}})$ satisfies,*

(a) $P_{\nu_{\mathcal{L},n}(\widetilde{\mathbf{M}})}(B_\epsilon(\widetilde{\mathbf{M}}^0)) \to 1$.

(b) $\liminf_n \inf_{\widetilde{\mathbf{M}} \notin B_\epsilon(\widetilde{\mathbf{M}}^0)} \ell_{\mathcal{L},n}/n > E_{\mathcal{L}}(D_{KL}(\mathbf{a}||\mathbf{M}^{0\prime}\mathbf{b}))$.

Assumption 1 is needed to ensure that the loss-functions $\ell_{\mathcal{L},n}$ are convex ensuring direct applicability of the results from Miller, 2019 to prove the lemma. To interpret Assumption 1, we consider the special case where we observe the true labels y, and the predicted labels $\mathbf{a}$ are categorical. Then this condition reduces to the statement that for every (i, j) pair, there are cases in $\mathcal{L}$ for whom the true class is i and the predicted class is j. This is of course necessary to estimate the misclassification rate M_{ij}. Thus, Assumption 1 can be interpreted as the limited labeled test set having data enough correctly estimate the sensitivities and specificities of the classifier for all class-pairs.

Lemma 1(a) is important on its own right as it establishes an important consistency result for model-free Bayesian estimating equations for compositional

regression. It states that when only loss $\ell_{\mathcal{L},n}$ is considered, the coefficients $\mathbf{M}$ for the regression equation (4.6) is consistently estimated by generalized posteriors from KLD loss function. We do not even need to actually observe the true labels y as observing the beliefs $\mathbf{b}$ suffices.

For our quantification problem, Lemma 1(b) is, however, the more relevant. As our f_N's are not convex, an alternate sufficient condition of Miller, 2019 to establish posterior concentration is that the infimum of f_N outside any neighborhood around the true $\boldsymbol{\theta}^0$ is strictly greater than $f(\boldsymbol{\theta}^0)$ for large enough N. Lemma 1(b) states that outside of any neighborhood around the true value $\mathbf{M}^0$, the empirical loss-function $\ell_{\mathcal{L},n}/n$ has higher value than the limiting loss-function $E_{\mathcal{L}}(D_{KL}(\mathbf{a}||\mathbf{M}'\mathbf{b}))$. A complementary result to Lemma 1(b) is that

Lemma 2. $\liminf_N \inf_{\boldsymbol{\theta}\in\Theta} \ell_{\mathcal{U},N}/N \geq E_{\mathcal{U}}(D_{KL}(\mathbf{a}||\mathbf{M}^{0'}\mathbf{p}^0))$.

Lemma 2 states that the infimum value of $\ell_{\mathcal{U},N}(\mathbf{M}'\mathbf{p})$ over the entire space Θ is greater than or equals to the limiting loss-function $E_{\mathcal{U}}(D_{KL}(\mathbf{a}||\mathbf{M}'\mathbf{p}))$ evaluated at true $\boldsymbol{\theta}^0$. Combining, Lemmas 1(b) and 2, we have that for any region R of Θ, $f_N(\boldsymbol{\theta})$ is greater than $f(\boldsymbol{\theta}^0)$ unless R lies in an infinitesimally small neighborhood around $\widetilde{\mathbf{M}}^0$. Thus, use of the local labeled set $\mathcal{L}$ via the loss function $\ell_{\mathcal{L},n}$ helps to identify $\mathbf{M}$, as the posterior is guaranteed to concentrate around $\mathbf{M}^0$. As $\mathbf{M}$ concentrates around $\mathbf{M}^0$, the loss $\ell_{\mathcal{U},N}(\mathbf{M},\mathbf{p})$ becomes capable of identifying $\mathbf{p}$. A sufficient condition for this is that $\ell_{\mathcal{U},N}(\mathbf{M}^0,\mathbf{p})$ is a convex function of $\mathbf{p}$. Assumption 2 ensures this convexity. It is a *separability assumption* necessary for quantification as if there exists two probability vectors $\mathbf{p}^0$ and $\mathbf{p}^1$ such that $\mathbf{M}^{0'}\mathbf{p}^0 = \mathbf{M}^{0'}\mathbf{p}^1$ then it will be impossible to identify $\mathbf{p}$

based on predicted labels. This separability, or identifiability, assumption has long been discussed in the finite mixture model literature (Teicher, 1963; Yakowitz and Spragins, 1968), but has not been discussed for methods which rely on class-conditional first moments of classifier output for quantification.

Theorem 4 guarantees posterior concentration when using the actual generalized posterior ν. However, our Gibbs sampler relies on rounding and coarsening ν using an integer factor T. The following result connects the theory to the practical implementation.

Corollary 1. *Let $\nu_{coarse,N}$ denote the rounded and coarsened generalized posterior using a factor T_N with $T_N \to \infty$. Then, under the conditions for Theorem 4, we have $P_{\nu_{coarse,N}(\boldsymbol{\theta})}(B_\epsilon(\boldsymbol{\theta}^0)) \to 1$.*

Corollary 1 makes it evident that not only the coarsening step is important, the coarsening and rounding factor T_N needs to increase with increase of sample size.

Finally, it is trivial to extend the posterior concentration results for the ensemble quantification.

Corollary 2. *If K predictions are available for each label, and assumptions 1 and 2 are satisfied for each of the K prediction algorithms, then with $\boldsymbol{\theta} = (\widetilde{\mathbf{M}}^{(1)}, \ldots, \widetilde{\mathbf{M}}^{(K)}, \widetilde{\mathbf{p}})$ we have $P_{\nu_{coarse,N}(\boldsymbol{\theta})}(B_\epsilon(\boldsymbol{\theta}^0)) \to 1$.*

4.5 Simulations

We conduct multiple simulation studies to assess

1. accuracy of GBQL in estimating $\mathbf{p}$ in the presence of moderate amounts of labeled data
2. comparison of our estimating equations based approach with Dirchlet model-based approach using different data generating mechanisms
3. computation efficiency compared to Dirichlet model based approaches
4. estimation accuracy when there is uncertainty for some true labels in $\mathcal{L}$.

To mimic the motivating verbal autopsy situation, we used $N = 1000$, $n = 300$, $C = 5$, $\mathbf{p}_{\mathcal{L}} = E_{\mathcal{L}}(y_r) = (\frac{1}{C}, \ldots, \frac{1}{C})$, and the following four different values of $\mathbf{p}$ representing each of the four countries in the PHMRC dataset (Section 4.1)

$$\mathbf{p1} = (.20, .19, .27, .27, .07)$$

$$\mathbf{p2} = (.11, .11, .40, .29, .09)$$

$$\mathbf{p3} = (.09, .18, .52, .19, .02)$$

$$\mathbf{p4} = (.13, .30, .35, .19, .03)$$

We generated true labels

$$y_r|\mathbf{p}, \mathbf{p}_{\mathcal{L}} \sim \begin{cases} Multinomial(1, \mathbf{p}), \ r \in \mathcal{U} \\ Multinomial(1, \mathbf{p}_{\mathcal{L}}), \ r \in \mathcal{L} \end{cases}$$

And first allow for full knowledge of these labels for $r \in \mathcal{L}$, which means that $\mathbf{b}_r|y_r = i$ equals $\mathbf{e}_i$ for $r \in \mathcal{L}$.

We then simulated outputs $\mathbf{a}_r|y_r$ directly from a model, so that we know the true data generating mechanism of the dataset shift. We let

$$\mathbf{M} = \begin{bmatrix} 0.65 & 0.35 & 0 & 0 & 0 \\ 0 & 0.35 & 0.65 & 0 & 0 \\ 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0 & 0 & 0 & 0.8 & 0.2 \\ 0 & 0.4 & 0 & 0 & 0.6 \end{bmatrix}$$

We used two data generating mechanisms for $\mathbf{a}_r|y_r$. The first mechanism corresponds to a zero-inflated Dirichlet mixture model:

$$a^*_{rj}|y_r = i, \mathbf{M}_{i*} \sim \begin{cases} 0, \text{ if } M_{ij} = 0 \\ Gamma(5M_{ij}, 1), \text{ else} \end{cases} \qquad j = 1, \ldots, C$$

$$a_{rj} = \frac{a^*_{rj}}{\sum_{k=1}^{C} a^*_{rk}}$$

The second data generating mechanism introduced overdispersion in the data:

$$\tau_r \sim .5 \cdot Uniform(.1, 1) + .5 \cdot Uniform(10, 20)$$

$$a^*_{rj}|y_r = i, \mathbf{M}_{i*} \sim \begin{cases} 0, \text{ if } M_{ij} = 0 \\ Gamma(\tau_r \cdot M_{ij}, 1), \text{ else} \end{cases} \qquad j = 1, \ldots, C$$

$$a_{rj} = \frac{a^*_{rj}}{\sum_{k=1}^{C} a^*_{rk}}$$

Instances for which $\tau_r \leq 1$ will have responses a_{rj} close to 0 and 1, while instances with larger values of τ_r will have a_{rj} clustered closer to the non-zero entries of $\mathbf{M}$.

We then compare our method's estimates of $\mathbf{p}$ with estimates from following standard Bayesian Dirichlet mixture model which assumes the first data generating mechanism as truth.

$$y_r|\mathbf{p} \sim Multinomial(1, \mathbf{p})$$

$$\mathbf{a}_r|y_r = i \sim Dirichlet(\tau_i \cdot \mathbf{M}_{i*})$$

$$\tau_i \sim Normal(0, 25)$$

For both the Dirichlet model and the GBQL model, we used Dirichlet priors for $\mathbf{M}$ shrinking towards $\mathbf{I}$, and uninformative Dirichlet prior for $\mathbf{p}$.

Since the Dirichlet distribution does not support zeros, for running the Dirichlet model, 0 values were replaced with $\epsilon = .001$ and each $\mathbf{a}_r$ was renormalized. Posterior sampling for this model was performed using RStan Version 2.19.2 (Stan Development Team, 2019). Note that this model then becomes misspecified for the second true data generating mechanism. For both models, we ran three chains each with a total of 6,000 draws and a burn-in of 1,000 draws. We used the posterior mean of $\mathbf{p}$ as $\hat{\mathbf{p}}$.

To compare estimates of $\mathbf{p}$, we use a chance corrected version of the normalized absolute accuracy (NAA) (Gao and Sebastiani 2016). NAA is defined as

$$1 - \frac{\sum_{i=1}^{C} |p_i - \hat{p}_i|}{2(1 - min_i\{p_i\})}.$$

To represent random guessing of $\mathbf{p}$ with a score of 0, and perfect estimation of $\mathbf{p}$ with a score of 1, we follow Flaxman et. al (2015) and use the Chance

Corrected NAA

$$CCNAA = (NAA - .632)/(1 - .632).$$

We repeat our simulations 500 times for each choice of $\mathbf{p}$ and show the average CCNAA across this simulations in Figure 2. For case 1 (left panel) when the likelihood is correctly specified for the Dirichlet model, both methods produce accurate estimates of $\mathbf{p}$ and have approximately the same CCNAA. When we introduce overdispersion to the distribution of the $\mathbf{a}_r|y_r = i$ (right panel), we see that the performance the GBQL model is hardly affected, and substantially outperforms the now misspecified Dirichlet model in all cases.

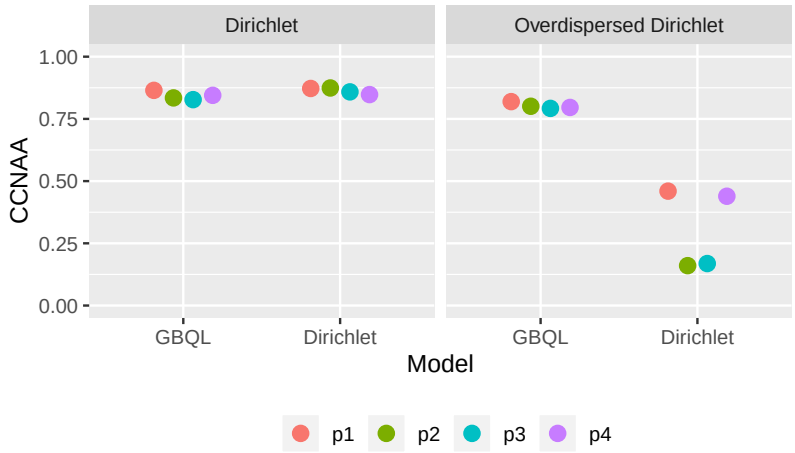


Figure 4.3: Columns shows results for the two different data generating mechanisms, while each color represents each of the four true values of p. The GBQL model produces high values of CCNAA for each of the scenarios, while assuming a Dirichlet mixture model likelihood only produces acceptable estimates of $\mathbf{p}$ when the likelihood correctly identifies the true data generating mechanism.

When we investigated the Stan output for the Dirichlet models, many of the chains failed to converge when the likelihood was misspecified (Table 4.1). Furthermore, on average the Stan model took nearly 200 times longer

to run than the GBQL method (Table 4.1). Thus, GBQL accurately estimates $\mathbf{p}$, removes the need to correctly specify the likelihood, is fast, and does not require fine-tuning for the posterior samples to converge.

Value for $\mathbf{p}$	Average $\hat{R}$ GBQL	Average $\hat{R}$ Dirichlet	Average Runtime (minutes) GBQL	Average Runtime (minutes) Dirichlet
p1	1.03	3.32	0.15	29.79
p2	1.02	3.43	0.16	29.70
p3	1.03	3.12	0.16	28.84
p4	1.03	3.46	0.15	29.88

Table 4.1: Average $\hat{R}$, as a measure of posterior sampling convergence, and runtime in minute for each value of $\mathbf{p}$ was computed for when there is overdispersion in the data generating mechanism.

We now examine the behavior of the GBQL model in the case of uncertain labels. To induce this uncertainty, we generate the compositional $\mathbf{b}_r$ from the following overdispersed Dirichlet distribution

$$\tau_r \sim .5 \cdot Uniform(.1, 1) + .5 \cdot Uniform(10, 20)$$

$$\mathbf{b}_r \sim \begin{cases} Dirichlet(\tau_r \mathbf{p}), \ r \in \mathcal{U} \\ Dirichlet(\tau_r \mathbf{p}_{\mathcal{L}}), \ r \in \mathcal{L} \end{cases}$$

and generate $y_r | \mathbf{b}_r \sim Multinomial(1, \mathbf{b}_r)$. The data generating process for the $\mathbf{a}_r$ is the same as in the simulations with known labels. The compositional $\mathbf{b}_r$ are used as the uncertain labels for $r \in \mathcal{L}$. Figure 4.4 plots the average CCNAA from GBQL with known labels y against CCNAA of GBQL with unknown labels $\mathbf{b}$ for each value of $\mathbf{p}$ and data generating mechanism. It can be seen that introducing uncertainty in the labels results in slightly lower, but nearly identical performance for estimating $\mathbf{p}$.

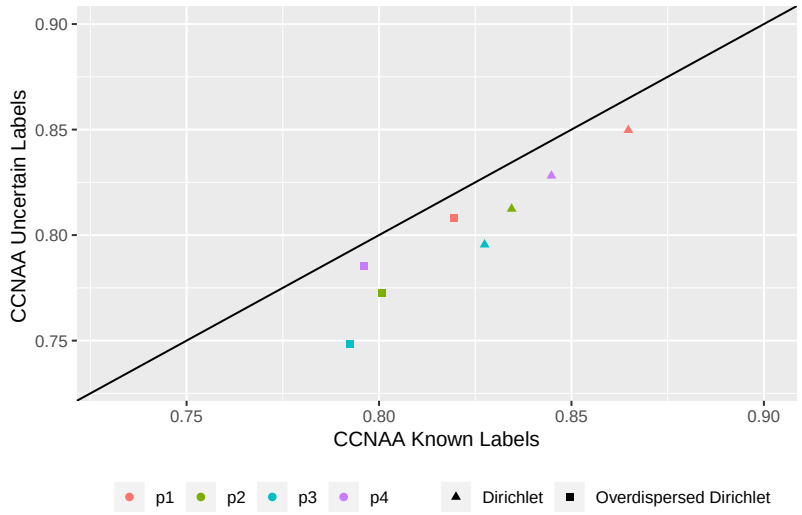
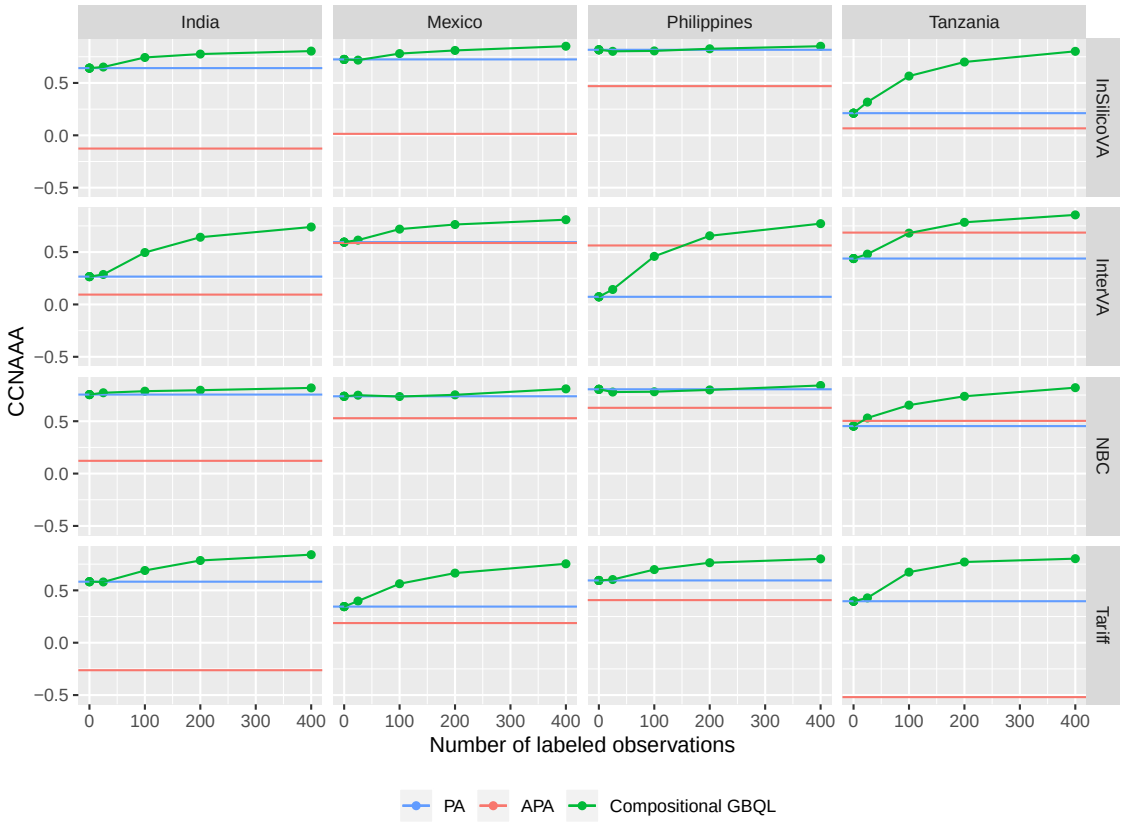


Figure 4.4: CCNAA for known versus uncertain labels using GBQL. Each color represents a different value for **p**, while the shapes represent the two different data generating mechanisms.

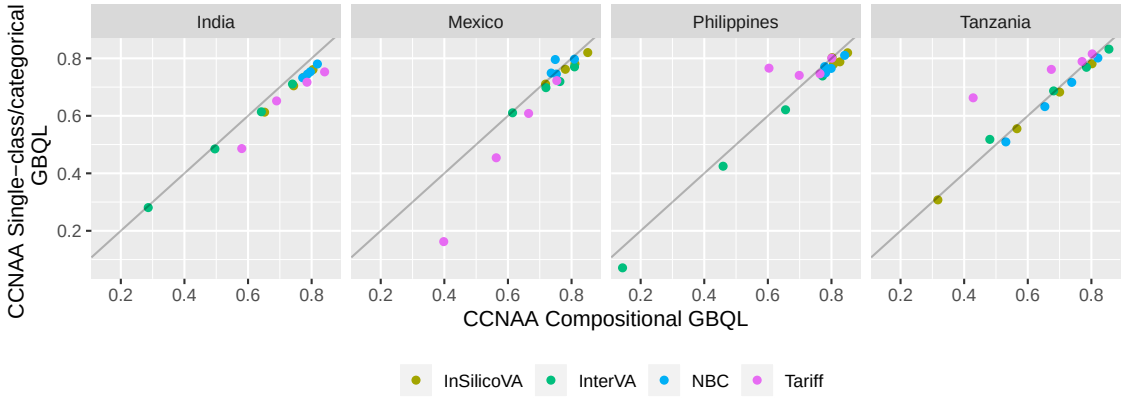
4.6 PHMRC Dataset Analysis

We now apply GBQL to the PHMRC dataset introduced in Section 1. The number of observations within India, Mexico, Philippines, and Tanzania are 2973, 1586, 1259, and 2023, respectively. To address country-specific dataset shift, for each country, we used the three remaining countries as training data for four methods commonly used for cause of death predictions: InterVA (Byass et al., 2012), InSilicoVA (McCormick et al., 2016), NBC (Miasnikof et al., 2015), and Tariff (Serina et al., 2015). The first three methods are probabilistic, while Tariff produces a score for each cause that needed to be normalized to be in $[0, 1]$. Model training was done using the openVA package version 1.0.8 (Li, McCormick, and Clark, 2019). We considered both compositional predictions

(for Tariff, this was the normalized score) and classifications (single-class categorical predictions based on the most likely cause of death for an individual per each algorithm). For comparisons, we obtained estimates using the CC and PA estimates of $\mathbf{p}$, that should align with the GBQL estimate for $n = 0$ (Section 4.3.6), as well as estimates using the ACC and APA methods. For GBQL in the country not used in training data, we then sampled labeled data of varying sizes (n=25, 100, 200, 400) to investigate the effect of increasing the number of known labels. Sampling was performed such that $\mathbf{p}_{\mathcal{L}} = (\frac{1}{5}, \ldots, \frac{1}{5})$, as in Section 4.5. We repeated this 500 times for each size of n. Results for the average CCNAA when using compositional predictions are shown in Figure 4.5a, while Figure 4.5b compares the CCNAA for GBQL using compositional predictions versus GBQL using single-class categorical predictions.



(a) Average CCNAA for increasing numbers of labeled observations across all countries in the PHMRC dataset for four common VA algorithms. Average CCNAA for GBQL using compositional predictions is shown in green. We also compare the performance of GBQL to the PA (blue) and APA (red) methods



(b) Comparison of CCNAA between GBQL using compositional predictions versus single-class/categorical predictions. Each point represents a different value of n, with the black line representing the identity line.

Figure 4.5: GBQL outperforms PA and APA for PHMRC quantification, while handling both compositional and single-class predictions

When no labeled instances are available, we see that the APA method performs worse than the PA method across almost all countries and algorithms, demonstrating why it is not appropriate to estimate $\mathbf{M}$ using the training data in the presence of dataset shift. We see that obtaining $n = 25$ labeled instances (an average of only 5 labeled deaths per class) does not effectuate any improvement in the performance over not having any labeled test data ($n = 0$). However, increasing this to 100 labels (an average of 20 labeled deaths per class) leads to large increase in CCNAA indicating substantial improvement in estimation of $\mathbf{p}$ across all countries and algorithms. As there are 168 covariates used for building these classifiers, using just 100 observations to build a reliable classifier would be difficult, if not impossible. Quantification accuracy continues to increase with a larger number of labeled observations across all countries and algorithms, although the extent of this improvement is quite variable. Finally, comparing the performance of GBQL using the categorical versus compositional predictions, Figure 4.5b shows that overall, using compositional data offers slight improvement.

Figure 4.5a shows that classifier performance varies widely across settings. We now look at the performance of our ensemble method which uses predictions from all four algorithms. Figure 4.6 shows the CCNAA for the ensemble method and the individual algorithms for different numbers of labeled observations and each country. With only 25 labeled observations, the ensemble CCNAA is approximately an average of the CCNAA for each of the other algorithms, which is what we would expect, as for $n = 0$ it is exactly the average as discussed in Section 4.3.6. With more labeled observations, the

ensemble begins to either outperform all of the methods, or has CCNAA very close to that of the top performing method. Importantly, the ensemble method significantly outperforms the worst method for all combinations of country, output format and numbers of labeled observations, showing that including multiple algorithms and using the ensemble quantification protects against inadvertently selecting the worst algorithm.

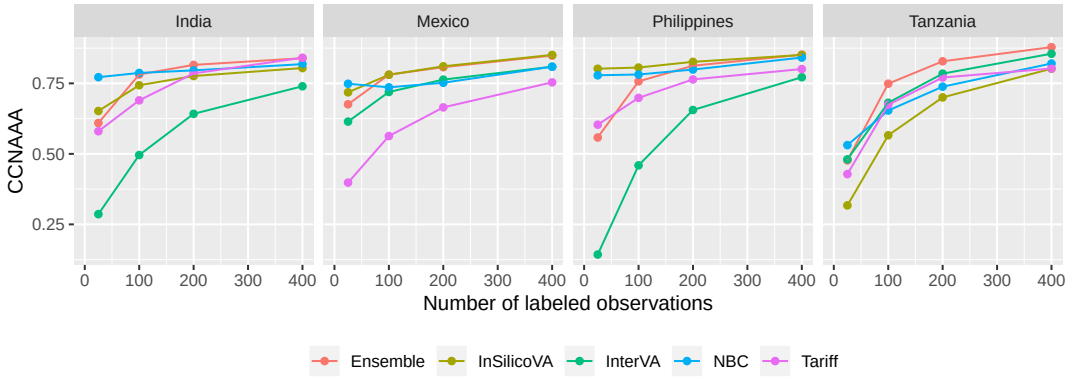


Figure 4.6: CCNAA comparing the ensemble GBQL (red) with the 4 individual GBQL algorithms across countries for both classification predictions and probalistic predictions

Finally, to illustrate the efficacy of GBQL even when true labels are observed with uncertainty, we create a toy dataset by randomly pairing individuals within a country in the PHMRC data. To introduce label uncertainty into the analysis, for a pair of individuals, $r1$ and $r2$, we let

$$b_{r1i} = b_{r2i} = \frac{1}{2}(I(y_{r1i} = 1) + I(y_{r2i} = 1)),$$

By using two individuals each with a single true label, we create two individuals each with uncertain true labels in such a way that the total number

of individuals with a given cause remains same in this new dataset as that in the actual PHMRC dataset. In other words, the data generation satisfies the assumption that $p(y_r = i|b_{ri}) = b_{ri}$. We then used these beliefs instead of the true labels as input for our method. Figure 4.7 compares the CCNAA for the individual methods across each value of n for compositional predictions when using the known labels versus representing uncertainty in the labels through beliefs, and shows that the performance of our method is nearly identical for both inputs.

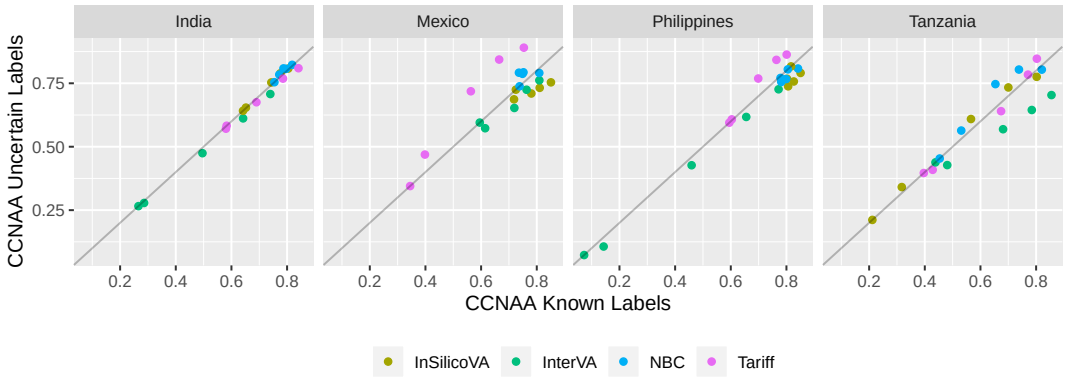


Figure 4.7: Comparison of CCNAA when using known labels versus labels with uncertainty. Each point represents a different value of n, with the black line representing the identity line.

4.7 Discussion

Quantification is an important and challenging problem that has only recently gained the attention it deserves. There are important limitations of the commonly used methods; CC (Forman, 2005), ACC (Forman, 2005), PA (Bella et al., 2010), and APA (Bella et al., 2010) do not use a probabilistic framework and only use training data, and therefore do not account for dataset shift, while

BTL (Datta et al., 2018) does not allow uncertainty in either the predicted or the true labels. The GBQL approach discussed here is more general, allowing for both categorical and compositional classifier output, incorporation of training data (through priors) and labeled data, and uncertain knowledge of labeled data classes. Thus, the GBQL model subsumes and extends these current methods.

By incorporating the categorical and compositional classifier output through the KLD loss function, which only relies on a simple first-moment assumption that is coherent for both data types, we circumvent the need for full model specification. In addition, the GBQL loss function is easily extended to harmonize output from multiple classifiers, leading to a unified ensemble method. Our application of the results of Bissiri, Holmes, and Walker, 2016 allows for model-free Bayesian inference, which in turn enables use of shrinkage priors to inform the estimation of $\mathbf{M}$ and $\mathbf{p}$ when no or limited labeled data from the test set is available. The GBQL generalized Gibbs posterior exhibits posterior consistency, as does the coarsened posterior used for extremely fast posterior sampling. Finally, extensive simulations and PHMRC data analysis show that the GBQL model is robust to model misspecification, and uncertainty in true labels, and significantly improves quantification in the presence of dataset shift.

Currently the GBQL method gives equal weight to all instances in $\mathcal{U}$ and $\mathcal{L}$. For ongoing quantification projects, $p_{test}(\mathbf{x}, y)$ may not be stable over time, and equally weighting instances collected early during the project may lead to inaccurate estimates of the current value for $\mathbf{p}$. A potential solution

could incorporate power priors (Ibrahim et al., 2015) for earlier observations, although we leave this for future research.

Further research is more warranted on general moment-based Bayesian methods for compositional data that builds on our application of the results from Bissiri et al. (2016). Given that our loss function is the one used in MQL based regression approaches, this justifies using priors on the regression parameters of interest, and updating these beliefs with a MCMC based method. In addition, our method could generally be used for semi-supervised mixture modeling of compositional observations. Important future contributions would be instance level class predictions and incorporation of higher moments through our loss function approach.

References

Moons, Karel GM, Andre Pascal Kengne, Diederick E Grobbee, Patrick Royston, Yvonne Vergouwe, Douglas G Altman, and Mark Woodward (2012). "Risk prediction models: II. External validation, model updating, and impact assessment". In: *Heart* 98.9, pp. 691–698.

Giachanou, Anastasia and Fabio Crestani (2016). "Like it or not: A survey of twitter sentiment analysis methods". In: *ACM Computing Surveys (CSUR)* 49.2, p. 28.

Valdez, Ashley, Elizabeth Ellen Hancock, Seyi Adebayo, David Kiernicki, Daniel Proskauer, John R Attewell, Lucinda Bateman, Alfred DeMaria Jr, Charles Warren Lapp, Peter C Rowe, et al. (2018). "Estimating Prevalence, Demographics and Costs of ME/CFS Using Large Scale Medical Claims Data and Machine Learning". In: *Frontiers in pediatrics* 6, p. 412.

King, Gary, Ying Lu, et al. (2008). "Verbal autopsy methods with multiple causes of death". In: *Statistical Science* 23.1, pp. 78–91.

McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark (2016). "Probabilistic cause-of-death assignment using verbal autopsies". In: *Journal of the American Statistical Association* 111.515, pp. 1036–1049.

Serina, Peter, Ian Riley, Andrea Stewart, Spencer L James, Abraham D Flaxman, Rafael Lozano, Bernardo Hernandez, Meghan D Mooney, Richard Luning, Robert Black, et al. (2015). "Improving performance of the Tariff Method for assigning causes of death to verbal autopsies". In: *BMC medicine* 13.1, p. 291.

Byass, Peter, Daniel Chandramohan, Samuel J Clark, Lucia D'ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. (2012). "Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool". In: *Global health action* 5.1, p. 19281.

Miasnikof, Pierre, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha (2015). "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths". In: *BMC medicine* 13.1, p. 286.

Forman, George (2005). "Counting positives accurately despite inaccurate classification". In: *European Conference on Machine Learning*. Springer, pp. 564–575.

Bella, Antonio, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana (2010). "Quantification via probability estimators". In: *2010 IEEE International Conference on Data Mining*. IEEE, pp. 737–742.

González, Pablo, Alberto Castaño, Nitesh V Chawla, and Juan José Del Coz (2017). "A review on quantification learning". In: *ACM Computing Surveys (CSUR)* 50.5, p. 74.

Pérez-Gállego, Pablo, Alberto Castano, José Ramón Quevedo, and Juan José del Coz (2019). "Dynamic ensemble selection for quantification tasks". In: *Information Fusion* 45, pp. 1–15.

Kalter, Henry D, Abdoulaye-Mamadou Roubanatou, Alain Koffi, and Robert E Black (2015). "Direct estimates of national neonatal and child cause–specific mortality proportions in Niger by expert algorithm and physician–coded analysis of verbal autopsy interviews". In: *Journal of global health* 5.1.

Forman, George (2008). "Quantifying counts and costs via classification". In: *Data Mining and Knowledge Discovery* 17.2, pp. 164–206.

Westreich, Daniel, Jessie K Edwards, Catherine R Lesko, Elizabeth Stuart, and Stephen R Cole (2017). "Transportability of trial results using inverse odds of sampling weights". In: *American journal of epidemiology* 186.8, pp. 1010–1014.

Cole, Stephen R and Elizabeth A Stuart (2010). "Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial". In: *American journal of epidemiology* 172.1, pp. 107–115.

Fawcett, Tom and Peter A Flach (2005). "A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions". In: *Machine Learning* 58.1, pp. 33–38.

Moreno-Torres, Jose G, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera (2012). "A unifying view on dataset shift in classification". In: *Pattern Recognition* 45.1, pp. 521–530.

Murray, Christopher JL, Alan D Lopez, Robert Black, Ramesh Ahuja, Said Mohd Ali, Abdullah Baqui, Lalit Dandona, Emily Dantzer, Vinita Das,

Usha Dhingra, et al. (2011). "Population Health Metrics Research Consortium gold standard verbal autopsy validation study: design, implementation, and development of analysis datasets". In: *Population health metrics* 9.1, p. 27.

Datta, Abhirup, Jacob Fiksel, Agbessi Amouzou, and Scott Zeger (2018). "Regularized Bayesian transfer learning for population level etiological distributions". In: *arXiv preprint arXiv:1810.10572*.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall. ISBN: 9780412317606. URL: http://books.google.com/books?id=h9kFH2_FfBkC.

Murphy, Kevin P et al. (2006). "Naive bayes classifiers". In: *University of British Columbia* 18, p. 60.

Specht, Donald F (1990). "Probabilistic neural networks". In: *Neural networks* 3.1, pp. 109–118.

Dahinden, Corinne (2011). "An improved Random Forests approach with application to the performance prediction challenge datasets". In: *Hands-on Pattern Recognition, Challenges in Machine Learning* 1, pp. 223–230.

Bragg, Jonathan, Daniel S Weld, et al. (2013). "Crowdsourcing multi-label classification for taxonomy creation". In: *First AAAI conference on human computation and crowdsourcing*.

Szczurek, Ewa, Przemysław Biecek, Jerzy Tiuryn, and Martin Vingron (2010). "Introducing knowledge into differential expression analysis". In: *Journal of Computational Biology* 17.8, pp. 953–967.

Quevedo, José RamóN, Oscar Luaces, and Antonio Bahamonde (2012). "Multilabel classifiers with a probabilistic thresholding strategy". In: *Pattern Recognition* 45.2, pp. 876–883.

Bissiri, Pier Giovanni, Chris C Holmes, and Stephen G Walker (2016). "A general framework for updating belief distributions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 1103–1130.

Hopkins, Daniel J and Gary King (2010). "A method of automated nonparametric content analysis for social science". In: *American Journal of Political Science* 54.1, pp. 229–247.

Hijazi, Rafiq H and Robert W Jernigan (2009). "Modelling compositional data using Dirichlet regression models". In: *Journal of Applied Probability & Statistics* 4.1, pp. 77–91.

Wong, Tzu-Tsung (1998). "Generalized Dirichlet distribution in Bayesian analysis". In: *Applied Mathematics and Computation* 97.2-3, pp. 165–181.

Tang, Zheng-Zheng and Guanhua Chen (2018). "Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis". In: *Biostatistics*.

Comas-Cufí, Marc, Josep Antoni Martín-Fernández, and Glòria Mateu-Figueras (2016). "Log-ratio methods in mixture models for compositional data sets". In: *SORT-Statistics and Operations Research Transactions* 1.2, pp. 349–374.

Liang, Kung-Yee and Scott L Zeger (1986). "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1, pp. 13–22.

Papke, Leslie E and Jeffrey M Wooldridge (1996). "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates". In: *Journal of applied econometrics* 11.6, pp. 619–632.

Mullahy, John (2015). "Multivariate fractional regression estimation of econometric share models". In: *Journal of Econometric Methods* 4.1, pp. 71–100.

Bhattacharya, Anirban, Debdeep Pati, Yun Yang, et al. (2019). "Bayesian fractional posteriors". In: *The Annals of Statistics* 47.1, pp. 39–66.

Ibrahim, Joseph G, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen (2015). "The power prior: theory and applications". In: *Statistics in medicine* 34.28, pp. 3724–3749.

Miller, Jeffrey W and David B Dunson (2019). "Robust Bayesian inference via coarsening". In: *Journal of the American Statistical Association* 114.527, pp. 1113–1125.

Kessler, Lawrence M and Murat K Munkin (2015). "Bayesian estimation of panel data fractional response models with endogeneity: an application to standardized test rates". In: *Empirical Economics* 49.1, pp. 81–114.

Yuan, Z, R Chappell, and H Bailey (2007). "The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach". In: *Biometrics* 63.1, pp. 173–179.

Miller, Jeffrey W (2019). "Asymptotic normality, concentration, and coverage of generalized posteriors". In: *arXiv preprint arXiv:1907.09611*.

Teicher, Henry (1963). "Identifiability of finite mixtures". In: *The annals of Mathematical statistics*, pp. 1265–1269.

Yakowitz, Sidney J and John D Spragins (1968). "On the identifiability of finite mixtures". In: *The Annals of Mathematical Statistics*, pp. 209–214.

Stan Development Team (2019). *RStan: the R interface to Stan*. URL: http://mc-stan.org/.

Li, Zehang, Tyler McCormick, and Sam Clark (2019). *openVA: Automated Method for Verbal Autopsy*. URL: https://CRAN.R-project.org/package=openVA.

Chapter 5

Improving Verbal-Autopsy-based Cause Specific Mortality Fraction Estimates in Mozambique using Bayesian machine learning

5.1 Introduction

The Countrywide Mortality Surveillance for Action (COMSA)-Mozambique seeks to provide timely cause specific mortality fractions (CSMF) at a national level for the country of Mozambique. Many deaths which are registered as part of COMSA occur outside of a hospital and thus are not assigned an official cause of death (COD). To give informed CSMF estimates, COMSA performs a verbal autopsy (VA) for each registered death. Recent efforts have resulted in standard VA questionnaires that allow for informed CSMF estimates (Nichols et al., 2018).

For each VA, trained COMSA VA data collectors collect information on 354 symptoms. Standard practice for assigning a COD based on a VA is to have

two physicians review the VA (Soleman, Chandramohan, and Shibuya, 2006). However, this process is timely and costly, and would prevent COMSA from presenting up-to-date mortality statistics on a rolling basis.

Automated, computer-coded classifiers for VA algorithms (CCVA) such as InSilicoVA (McCormick et al., 2016), InterVA-4 (Byass et al., 2012), the Naives Bayes Classifier (NBC) for Verbal Autopsies (Miasnikof et al., 2015), and the expert algorithm for verbal autopsy (EAVA) (Kalter, Perin, and Black, 2016) allow for fast assignments of CODs from VA data. Aggregating these COD assignments to produce CSMF estimates from VA algorithms can produce results similar to that from physician review (Jha et al., 2019). However, the data or expert knowledge at the heart of these CCVA algorithms that relates VA responses to COD information has a large influence on VA algorithm accuracy (Clark, Li, and McCormick, 2018). Even for individuals who die from the same cause, VAs performed in Mozambique may result in different symptom information than VAs performed in another country, due to culture differences and other factors specific to the local context in Mozambique. Because the algorithms used in this study were developed without knowledge of the local context that relates VA symptom information to COD in Mozambique, these algorithms may be inaccurate when applied to the COMSA VA data (Clark, Li, and McCormick, 2018; Datta et al., 2020).

Previous work has shown how to improve CSMF estimates by collecting gold-standard COD (GS-COD) information on a small number of deaths for whom a VA has also been performed (Datta et al., 2020; Fiksel et al., 2020). This GS-COD information is used to learn the misclassification rates of the VA

algorithms, which is used to calibrate the original CSMF estimates by taking into account for the imperfect sensitivity and specificity of the CCVA classifier.

To correct for imperfect VA algorithm COD predictions for deaths in Mozambique, we use data from the Child Health and Mortality Prevention (CHAMPS) project. CHAMPS is an ongoing surveillance project that performs a minimally invasive autopsy (MIA), also known as a minimally invasive tissue sample (MITS) (Byass, 2016) to determine the COD with high precision. A VA for each death that occurs within a CHAMPS site (*CHAMPS Cause of Death Data*) is also conducted. MITS COD assignments have been shown to be very accurate compared to complete diagnostic autopsies (Castillo et al., 2016). However, because MITS COD assignments are decided on by an expert human panel, there may be some uncertainty in the final cause assignment.

We use the MITS and VA data collected on child (1-59 months old) deaths that occurred at the CHAMPS sites (including Mozambique) to estimate the misclassification rates of two VA algorithms InsilicoVA and EAVA. These estimates reveal substantial classification errors for both algorithms cautioning against the use of the raw CSMF estimates as they are likely to be very biased. We use the misclassification matrices to produce calibrated VA CSMF estimates for child deaths in Mozambique. We use the Generalized Bayesian Quantification Learning (GBQL) (Fiksel et al., 2020) framework to handle uncertainty in MITS COD classification, as well as to incorporate probabilistic individual COD predictions from VA algorithms. This framework also allows for a single CSMF estimate based on an ensemble learner that incorporates VA COD assignments from both InSilicoVA and EAVA, rather than having to

choose a CSMF from one of the two algorithms. We demonstrate a complete workflow of the methodology that first estimates the raw CSMF estimates and misclassification rates, combines them to produce calibrated CSMF estimates, and provides quantitative model comparison metrics to compare and choose between the raw and calibrated CSMF estimate.

5.2 Data

We use two main sources of data to estimate the CSMF for 1-59-month old children in Mozambique. The first source consists of VA data for 989 child deaths from the ongoing nationally representative VA survey conducted by COMSA. We refer to this source of data as the COMSA data. The second source of data is obtained from 283 child deaths that occurred within hospitals. Information about these deaths was collected by the CHAMPS project, and we refer to this source of data as the CHAMPS data. These deaths occurred in CHAMPS sites across several countries: South Africa (n = 115), Kenya (n=115), Mozambique (n=26), Mali (n=23), Ethiopia (n=3), and Bangladesh (n=1). For each of these deaths, the CHAMPS project performed both a VA and a MITS. For each MITS from the CHAMPS project, there is an underlying COD and an immediate COD, if applicable. Table 5.1 shows the number of deaths for each combination of underlying and immediate COD. For many cases, it is ambiguous which among the immediate and the underlying cause should be the GS-COD and we consider both causes in our methodology taking into account this uncertainty. For deaths for which the underlying cause is determined to be the sole COD, we do not consider a second cause.

		Immediate COD							
		Malaria	Pneumonia	Diarrhea	Severe Malnutrition	HIV	Other	Other Infections	Total
Underlying COD	Malaria	16	3	0	0	0	1	1	23
	Pneumonia	0	21	0	0	0	4	13	38
	Diarrhea	0	5	15	0	0	0	3	23
	Severe Malnutrition	4	8	2	0	0	0	23	37
	HIV	4	14	3	0	0	1	14	36
	Other	1	26	0	0	0	28	38	93
	Other Infections	0	44	0	0	0	7	22	33
	Total	25	81	20	0	0	41	116	283

Table 5.1: Classification of GS-COD in the CHAMPS dataset by underlying and immediate COD

5.3 Methods

5.3.1 Verbal Autopsy Algorithm Probabilities

We use both InSilicoVA and EAVA for individual COD assignments based off the VA symptom information for each death in the two sources of data. InSilicoVA is a probabilistic method, and gives individual probability estimates for 61 specific causes of death. To estimate these probabilities, InSilicoVA uses a conditional probability matrix which specifies the probability of observing each symptom, conditional on each COD. This matrix is derived from recommendations of an expert panel, and is the same one used for the InterVA algorithm (Byass et al., 2012). We aggregate these probabilities to the seven broad causes of death used in our study: pneumonia, malaria, diarrhea, severe malnutrition, HIV, other infections, and other cause of death. For each individual death, we thus get a 7×1 vector of estimated probabilities or percentages of the COD being in each of the seven broad categories.

EAVA uses an expert derived hierarchy to predict a COD from a VA. EAVA is deterministic and for each death may produce either a single most likely

COD, a first and second most likely COD, or conclude that the COD is unable to be determined. To handle the three potential outputs from EAVA, we map each output to a probabilistic interpretation. In the case of a single COD, this COD is assigned a probability of 100%. For a first and second most likely COD, the most likely COD is assigned a probability of 75%, and the second most likely COD is assigned a probability of 25%. Finally, for an undetermined COD, all causes are assigned equal probability.

We obtain "uncalibrated" CSMF estimates for each method by averaging the individual probability estimates over the 989 nationally representative deaths in the COMSA data. Formally, for both the InsilicoVA and EAVA algorithms we can write the predicted COD for the r^{th} case as a 7×1 vector a_r whose entries $a_{r1}, a_{r2}, \ldots, a_{r7}$ sum up to 1. For InsilicoVA, a_{rj} is going to be the estimated probability that the r^{th} individual died from cause j. For EAVA, only up to two of the a_{rj}'s are going to be non-zero for every individual. The uncalibrated estimate for each algorithm a is given by

$$\widehat{q}_a = \frac{1}{N} \sum_{r=1}^{N} a_r, \text{ for } a \in \{\text{EAVA, InsilicoVA}\}, N = 989. \tag{5.1}$$

Because it is impossible to know which method produces a more accurate CSMF estimate, we also produce an uncalibrated "ensemble" CSMF estimate (Fiksel et al., 2020), which is the average of the two methods' CSMF estimates.

5.3.2 Verbal Autopsy Algorithm Misclassification Rates

Misclassification occurs from a VA algorithm when a VA algorithm assigns an individual COD that is different from that individual's GS-COD. Both Chapter

2 and Chapter 4 of this thesis showed that estimating the misclassification rates of a VA algorithm to obtain a calibrated CSMF estimate can greatly improve over the uncalibrated CSMF estimate. For example, suppose there are only two causes of interest, and we know that a given CCVA has sensitivities for the two causes of 95% and 65%, respectively. This means that we expect the CCVA to predict a higher prevalence for the first cause than the true prevalence, as it mistakenly assigns 35% of people who truly die of the second cause to the first cause. After obtaining VA COD assignments for each 1000 individuals in our population set, we obtain an uncalibrated CSMF of 53% for the first cause and 47% for the second cause. However, calibrating this result with our known sensitivites gives us the correct calibrated CSMF of 30% and 70% for the two causes, respectively.

Because the misclassification rates for the CCVA algorithms are not known for our setting, we use the CHAMPS data to estimate these misclassification rates, considering the MITS COD as the GS-COD, and assuming that the misclassification rates in the COMSA and CHAMPS data are the same. If all the MITS cases were assigned a single COD (the underlying COD), and we only used the top cause for the VA (plurality-rule), then for each CHAMPS case we will have one MITS COD and one VA COD and the entries of the misclassification matrix $M = (M_{ij})$ can be estimated as:

$$M_{ij} = \frac{\text{\# of cases with MITS cause } i\text{, and VA cause } j}{\text{\# of cases with MITS cause } i} \tag{5.2}$$

However, in our case, two difficulties arise in deriving the misclassification

rate for a VA algorithm. The first difficulty is that we do not want to lose information by using the plurality-rule and select the top single-class predictions. Instead we adopt a fully probabilistic approach, using the full vector a_r of estimated probabilities from each algorithm. We resolve this by defining the misclassification "rate" for single-cause-MITS-multi-cause-VA as the average predicted probability for each cause, conditional on each gold-standard cause i.e.,

$$M_{ij} = \frac{\sum \text{ estimated VA probabilities } a_{rj} \text{ for cause } j, \text{ for cases with MITS cause } i}{\text{\# of cases with MITS cause } i}. \tag{5.3}$$

The definition in (5.3) is a generalization of (5.2) as they are same when all the VA's give a single COD.

The second difficulty is how to handle cases where both the MITS underlying and immediate causes are believed to be contributing to the death, meaning there is uncertainty in the GS-COD. Neither (5.2) or (5.3) applies to this situation as the phrase "cases with MITS cause i" used the numerator of both is no longer defined for MITS cases with more than one possible COD.

We adopt the same approach as for the cases with two possible COD predicted by EAVA. We translate this uncertain MITS output to a probabilistic interpretation. For the r^{th} case, we denote by g_r, a 7×1 vector encoding the MITS COD. For cases where MITS identifies a "single-cause", we give this underlying COD a probability of 100% and assign 1 to the corresponding element g_{rj} and 0's to the other components of g_r. For cases with the "multi-cause" interpretation, i.e., where both the immediate and underlying causes are believed to contribute to the outcome, we give both the underlying and

immediate COD a probability of 50%. This states that for death with different underlying and immediate causes, we believe both causes are equally likely to be the final COD, and translates to giving $1/2$ to the two corresponding entries fo g_r and 0's elsewhere.

For each CHAMPS case, we now have two 7×1 *compositional* vectors a_r and g_r representing the COD diagnosing from the VA and MITS respectively. We use the transformation-free method treating the probabilistic VA algorithm outputs a_r as the compositional outcome, and (possibly) multi-cause MITS probabilities g_r as the compositional predictor to define the misclassification rates for multi-cause-VA-multi-cause-MITS as

$$M = \underset{\{B=(B_{ij}):B_{ij}\geq 0,\sum_j B_{ij}=1\}}{\arg\min} \sum_{r=1}^{N} \text{kld}(a_r||B'g_r) \tag{5.4}$$

where $kld(y,x) = \sum_j y_j log(y_j/x_j)$ denotes the Kullback-Leibler Divergence loss function between two compositional vectors. Chapter 3 has demonstrated why for the special cases where only the VA is multi-cause, the estimate of M from (5.4) is identical to that from (5.3), which in turn is identical to (5.2) when the VA is also single-cause. Hence, the defnition in (5.4) is the most general one and we use it to estimate the multi-cause-VA-multi-cause-MITS misclassification rates for InSilicoVA and EAVA based on the CHAMPS data.

5.3.3 Bayesian calibration of VA and MITS COD Data

The uncalibrated CSMF estimate (5.1) from a VA algorithm essentially assumes that the VA algorithm has perfect sensitivity and specificity, i.e., the M matrix is the identity matrix. In practice, VA algorithms are prone to large

misclassification errors and the estimate of the misclassification matrix helps to quantify this. When a classifier demonstrates large misclassification rates, Chapter 2 outlines a frequentist two-step approach to calibrate CSMFs. This two stage approach first estimates the uncalibrated CSMF estimate and the misclassification matrices separately and combines them to yield the calibrated CSMF estimate. However, when the size of the dataset used to estimate the misclassification matrix is small this procedure is unstable. In our case, we have 283 CHAMPS cases to estimate a 7×7 misclassification matrix, leaving an average of 6 datapoints to estimates each entry of the matrix.

Chapter 2 also developed a joint estimation of the CSMF and the misclassification rates using a Bayesian approach. The joint model in the single-cause-VA-single-cause-MITS framework consists of two distinct pieces – a conditional multinomial model for the conditional probabilities (misclassification rates) of the VA COD given the MITS COD for the CHAMPS data, and a marginal multinomial model for the COMSA VA data with marginal probabilities derived from combining the misclassficiation rates and the true CSMF, i.e., the estimand of interest. In data-scarce settings such as ours, the Bayesian approach with “informative" priors substantially improves the calibration over the two-step frequentist approach by attaining a tradeoff between bias and variance. The informative prior ensures that in absence of enough paired MITS-VA data, the estimate of M is shrunk towards the identity matrix. This in turn shrinks the calibrated CSMF estimate towards the uncalibrated one which is the default estimate currently used by practitioners. In the extreme scenario, where there is no data to estimate the misclassification rates, the

calibrated and uncalibrated CSMF estimates are identical.

Chapter 4 extends the calibration method from Chapter 2 from a single-cause-VA-single-cause-MITS to our multi-cause-VA-multi-cause-MITS setup by switching from marginal or conditional probabilities to marginal or conditional averages, and replacing multinomial models with pseudo-multinomial likelihoods. This model thus incorporates the inherent uncertainty presented by the probabilistic COD information, rather than forcing practitioners to decide on the final COD, as when using the single-cause VA or MITS data. Furthermore, both Chapters 2 and 4 showed that the Bayesian hierarchical calibration model allows for an *ensemble* approach that helps circumvent the subjective decision-making about which VA algorithm to use to present final results. The ensemble method uses the entire suite of VA algorithms (in our case, InSilicoVA and EAVA) and models the misclassification rates from each of the algorithms, producing a single CSMF that is most coherent with all the misclassification rates. The ensemble calibration outperforms equal weighted equal weighting of the CSMFs from individual algorithms as it assigns weights in a data-driven manner generally ensuring higher weights for the better performing algorithm.

Because of the numerous advantages of the ensemble model with multi-cause GS-COD data, we use this model to obtain the estimate the CSMF for children of age 1-59 months in Mozambique. We outline the entire pipeline of obtaining the CSMF calibrated ensemble CSMF estimate using the COMSA VA using the CHAMPS VA-MITS data in Figure 5.1.

As a sensitivity analysis, we compare the ensemble multi-cause CSMF to

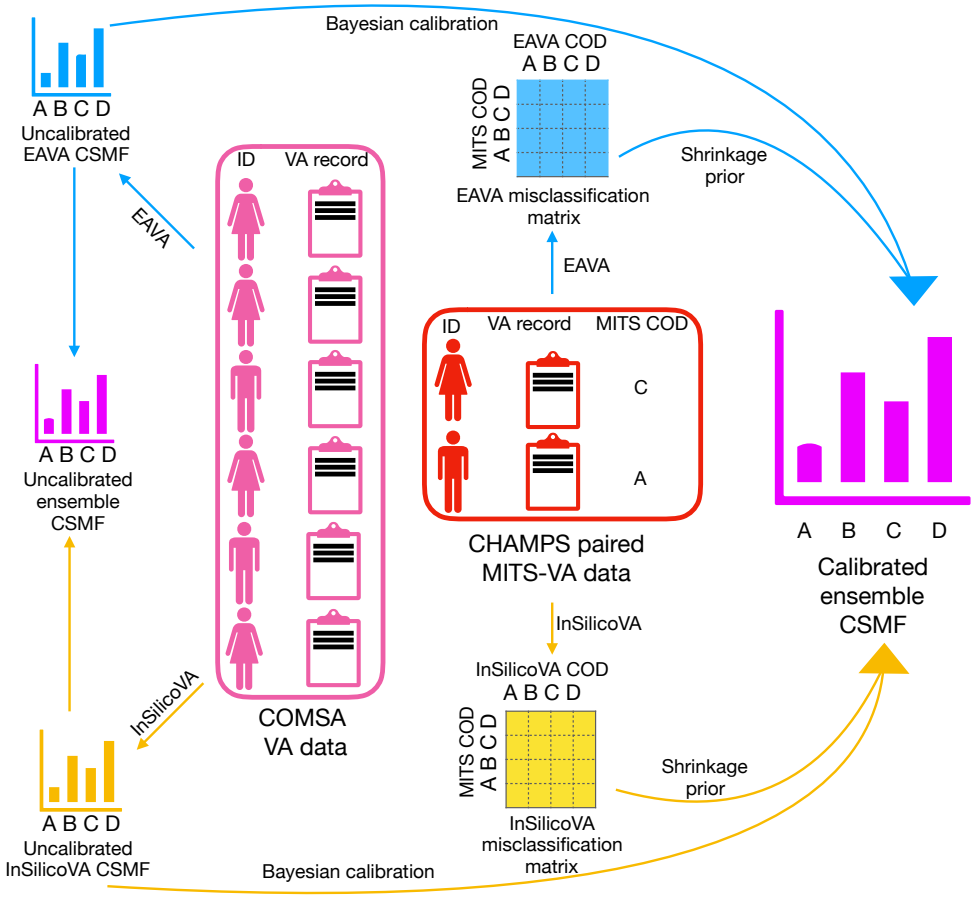


Figure 5.1: Pipeline for statistical calibration of CSMF estimates from a large VA data (COMSA VA data) using limited data with paired VA and true COD (CHAMPS MITS-VA data)

that obtained using the ensemble model but with the single-cause MITS COD (but with multi-cause VA), where we take the underlying COD to be the GS-COD. We also obtain calibrated CSMF estimates individually for InSilicoVA and EAVA, using both the single and multi-cause MITS data.

5.3.4 Model Selection and Comparison Using the WAIC

The Bayesian calibration model incorporates a shrinkage prior distribution on the misclassification rates that apriori posits that each algorithm has near-perfect sensitivity for every cause. Without the CHAMPS data, this prior means that the hierarchical calibration model would not change the uncalibrated CSMF estimates. This prior distribution necessitates choosing tuning parameter values which determine how strong this prior is, relative to the data. A strong prior would lead to estimating near-perfect sensitivity for a classifier leading to almost indistinguishable CSMF estimates before and after calibration. With a substantial number of gold-standard deaths this is undesirable. On the other hand, a very weak prior would lead to unstable estimates of the misclassification rates due to the extremely small size of the paired MITS-VA data leading which in turn destabilizes the final CSMF estimate.

To pick the tuning parameter values and also to compare the calibrated models to the uncalibrated models we use the widely applicable information criterion (WAIC) (Watanabe, 2010) which is an estimate of a model's ability to model future data, but using only already collected data (Vehtari, Gelman, and Gabry, 2017). In our case, we have two collected sources of data which are modeled differently in our model. The COMSA data correspond to a marginal multinomial (pseudo-)likelihood, and the CHAMPS data correspond to a conditional multinomial (pseudo-)likelihood. If we just used the COMSA data, which has no GS-COD information, to evaluate the WAIC, the best WAIC would be obtained by using the posterior distribution of the uncalibrated model for the CSMF. However, the uncalibrated CSMF assumes that the

models have perfect sensitivity and the CHAMPS data testifies for or against this assumption. Hence, the misclassification rates are critical to understand the relationship between the VA algorithm COD predictions and the GS-COD, and the WAIC calculation needs to include the CHAMPS data as well. Thus in our case, the future data for WAIC are twofold – VA algorithm COD predictions for nationally representative community deaths (COMSA data), which are modeled using both the true CSMF and misclassification rates, and VA algorithm COD predictions for the CHAMPS data with GS-COD, which are modeled using just the misclassification rates. If the CHAMPS data demonstrate large misclassification rates, the WAIC will be large thereby rightfully penalizing the uncalibrated CSMF for the wrong assumption of perfect sensitivity.

To estimate the WAIC from the calibrated models, we use the MCMC draws of the CSMF and the misclassification rates to obtain the posterior distribution of the loss-function presented in Chapter 4 for every death in both sources of data. We run the model for a grid of different combination of values of the tuning parameters and determine the optimal combination as one which minimizes the WAIC and ensures convergence of the MCMC to a unimodal marginal posterior distributions for each of the CSMFs and misclassification rates.

To estimate the WAIC for the uncalibrated models, we first obtain draws from the posterior distribution of the CSMF by assuming perfect sensitivity. As shown in Chapter 2, the posterior mean of this distribution is nearly exactly that of the uncalibrated CSMF estimate. The prefect sensitivity assumption

for the uncalibrated model, ideally translates to all posterior draws of the misclassification matrix being the identity matrix. This however produces a WAIC of $+\infty$ immediately ruling out the uncalibrated model. To make the WAIC of the uncalibrated model more competitive to that of the calibrated model, we compute the the former assuming model sensitivities of 95%, under the assumption that with sufficiently high sensitivity, one would still be willing to accept the uncalibrated CSMF estimates.

5.4 Results

5.4.1 Uncalibrated CSMFs:

We first present the uncalibrated CSMF estimates for InSilicoVA, EAVA, and the ensemble model in Figure 5.2 from COMSA. As the uncalibrated ensemble CSMF is simply the average of the two individual algorithm CSMFs, we focus on the differences between the InsilicoVA and EAVA estimates. Most notably, EAVA estimates a higher percentage of deaths from pneumonia (27% versus 19%), while InSilicoVA estimates a higher percentage of deaths from malaria (15% versus 6%). EAVA also estimates a higher percentage of deaths from severe malnutrition (10% versus 4%), and a slightly lower percentage of deaths from other causes of death (6% versus 10%). The two algorithms estimate similar percentages of death from the remaining causes.

5.4.2 Sensitivities and Misclassification rates:

Figure 5.3 shows the estimated uncalibrated misclassification matrices for InSilicoVA and EAVA, using both single-cause (5.3) and multi-cause (5.4)

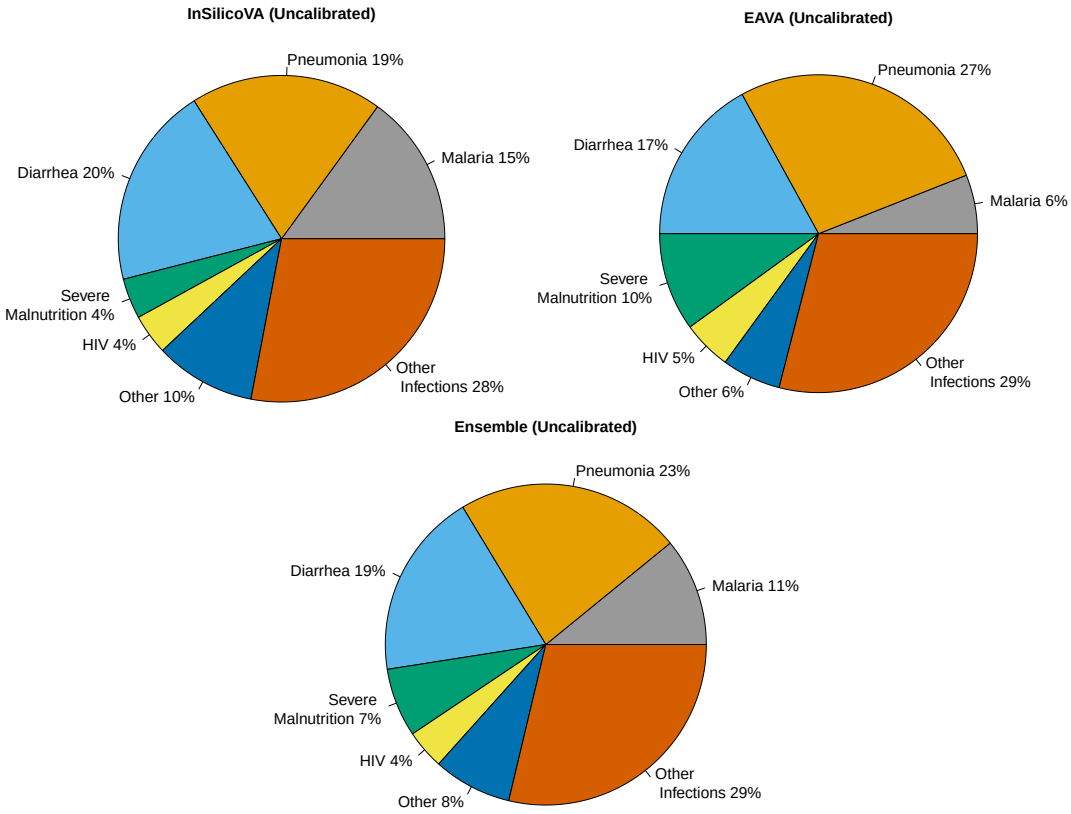


Figure 5.2: Uncalibrated CSMFs for InSilicoVA, EAVA, and the ensemble method.

MITS data from CHAMPS. Each point shows the expected probability a VA algorithm will predict for each cause on the columns, conditional on the GS-COD on the row. The algorithms show overall low estimated sensitivity (diagonal entries), with the EAVA sensitivity for diarrhea being the only sensitivity near 75%, and several of the cause-specific sensitivity rates being lower than 25%. InSilicoVA also shows very low sensitivities, with other infections and other COD being the only causes with a sensitivity above 25%. While Figure 5.2 shows that in the COMSA data both algorithms predict similar uncalibrated prevalence estimates for deaths due to diarrhea and other infections, Figure 5.3 suggests that based on the CHAMPS data the

misclassification rates conditional on the GS-COD being either of these two cause categories are very different between the two algorithms. For example, using the multi-cause MITS, EAVA has a sensitivity for diarrhea of 68%, while InSilicoVA has a sensitivity of 18%. However, InSilicoVA has a sensitivity for other infections of 57%, while EAVA has a sensitivity of 29%. Furthermore, among the CHAMPS cases, EAVA tends to predict higher probabilities for diarrhea and pneumonia being the COD, regardless of the GS-COD, while InSilicoVA tends to predict higher probabilities for other causes of death and other infections.

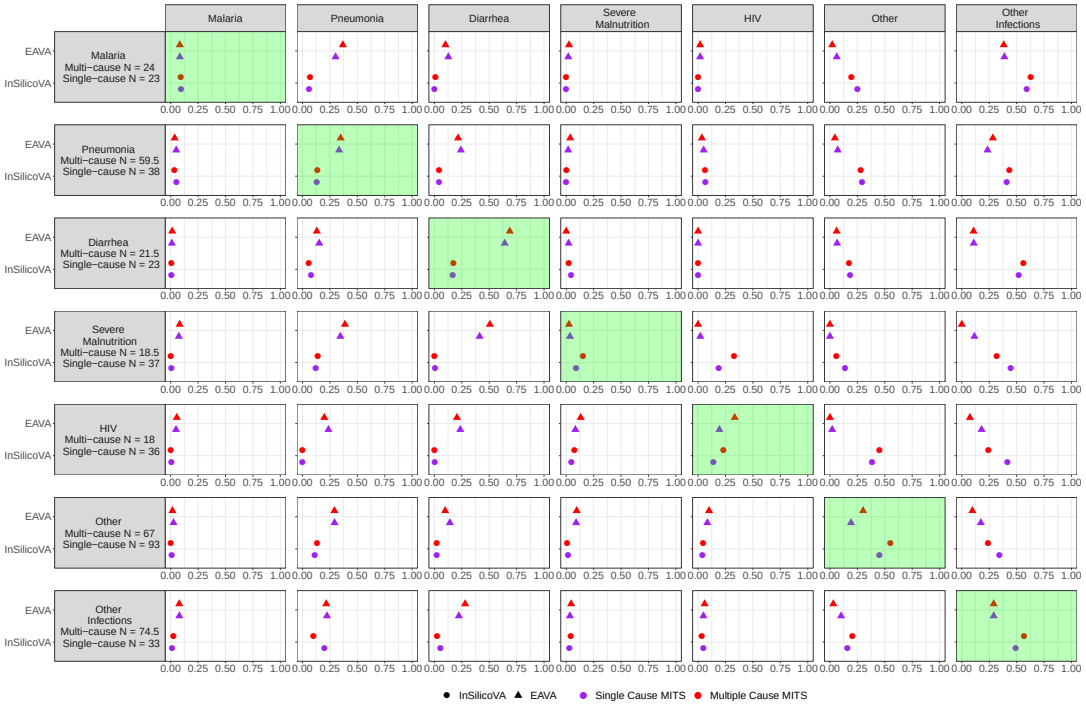


Figure 5.3: Uncalibrated misclassification rate estimates for EAVA (triangles) and InSilicoVA (circles), using both the multi-cause (red) and single-cause (purple) MITS data. The sample size for the multi-cause MITS is given by the sum of the individual GS-COD probabilities for each cause, while the sample size for the single-cause MITS is given by the number of individuals with the given cause as an underlying COD.

Figure 5.4 plots the estimated sensitivites for both algorithms when using the multi-cause MITS as opposed to the single-cause MITS. We see that using the multi-cause MITS COD data tends to result in slightly higher sensitivity estimates for each algorithm, as compared to using the single-cause data. This shows that by allowing for uncertainty in the GS-COD data the multi-cause calibration induced higher concordance between the VA- and MITS- predicted COD than the single-cause calibration.

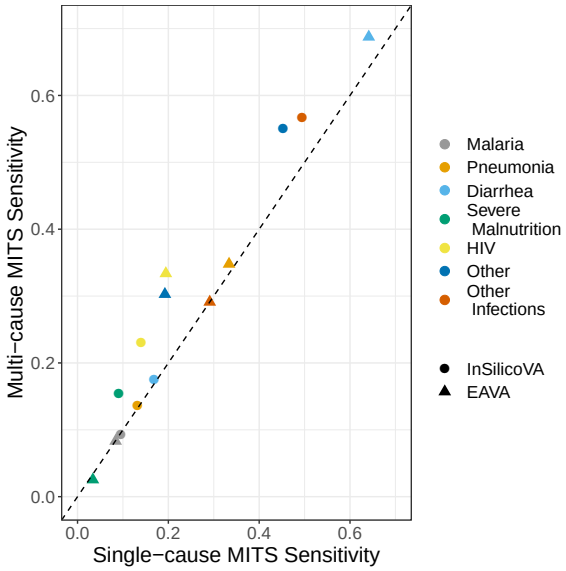


Figure 5.4: Cause-specific multi-cause MITS versus single-cause MITS estimated sensitivities for EAVA (triangles) and InSilicoVA (circles). The dashed line shows the identity line.

5.4.3 Calibrated CSMF estimates:

We present the final CSMF estimate using the ensemble calibration with multi-cause-VA-multi-cause-MITS COD data in Figure 5.5. The estimates indicate that pneumonia, other infections, diarrhea, and malaria are the most common causes of death, with mortality rates of 31%, 23%, 17%, and 16%, respectively.

Mortality fractions for severe malnutrition, HIV and other causes were all estimated to be less than 5%. Figure 5.6 compares the calibrated ensemble estimates with the uncalibrated ensemble estimates. The notable differences after the calibration are higher mortality rates are higher for pneumonia (31% from 23%) and malaria (16% from 11%), and lower for other infections (23% from 29%). In the supplemental figures section, we compare multi-cause-VA-multi-cause-MITS COD ensemble estimates are compared to the calibrated InSilicoVA and EAVA estimates.

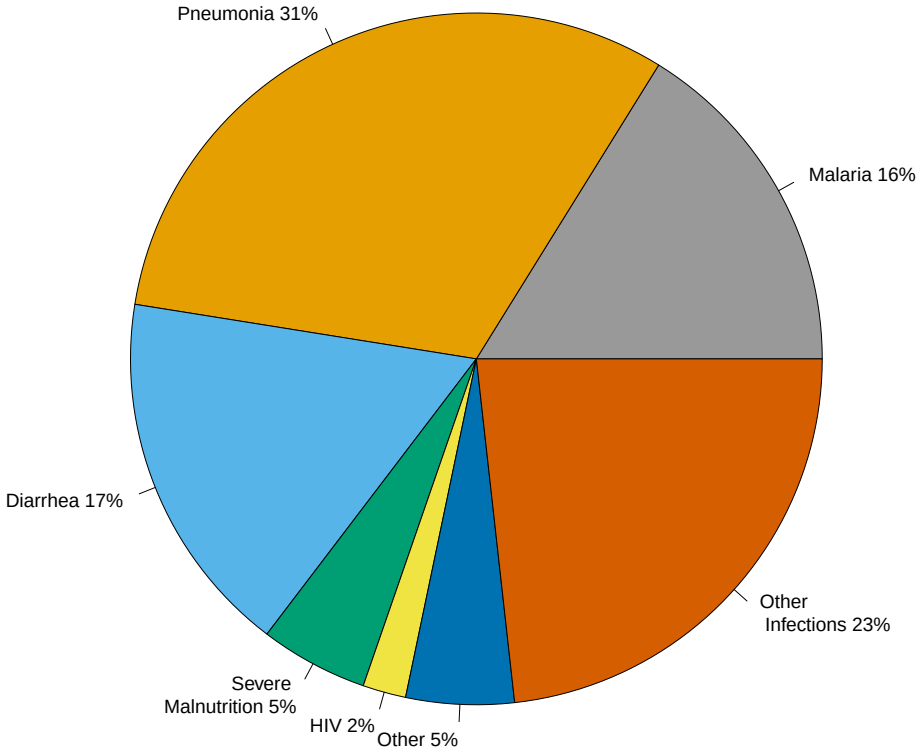


Figure 5.5: Calibrated CSMF estimates from the ensemble multi-cause MITS model.

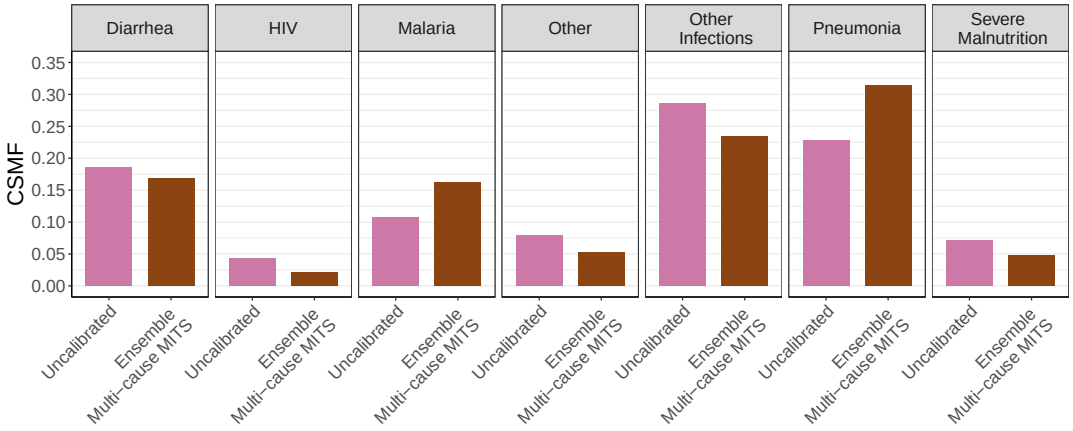


Figure 5.6: Comparison of CSMF estimates from the calibrated ensemble multi-cause MITS model versus the uncalibrated ensemble model.

5.4.4 Understanding the differences between calibrated and uncalibrated estimates:

To obtain some insights into the changes in CSMF after calibration, Figure 5.7 shows the estimates (posterior means) from the ensemble model of the algorithm-specific misclassification rates as compared to the estimated uncalibrated multi-cause misclassification rates (also presented in Figure 5.3). The calibrated sensitivities are significantly higher for both algorithms than the uncalibrated sensitivity estimates, across all causes. This is a result of a shrinkage prior on the misclassification rates, as well as low MITS sample sizes for several causes. The shrinkage for both algorithms is especially noticeable for diarrhea (N=21.5), severe malnutrition (N=18.5), and HIV (N=18), which are the least common multi-cause MITS causes.

The estimate for the EAVA malaria sensitivity from the ensemble calibration is below 40%. This is significantly lower than the sensitivity estimates for

the other causes across both algorithms. Low sensitivity means that EAVA often fails to diagnose malaria as the COD, despite it being the GS-COD, the ensemble model takes this into account by giving an increased estimate for the malaria mortality rate, relative to the uncalibrated estimate from Figure 5.2) (16% to 8%). The estimated malaria sensitivity for InSilicoVA is 75%, similar to the other estimated sensitivities, leading to the calibrated InSilicoVA estimate of the malaria mortality rate being similar to the calibrated InSilicoVA estimate from Figure 5.2) (16% versus 15%).

Figure 5.7: A comparison of the ensemble multi-cause MITS posterior mean misclassification rates (brown) to the uncalibrated misclassification rate estimates (pink) for InSilicoVA and EAVA. The sample size for the multi-cause MITS is given by the sum of the individual GS-COD probabilities for each cause.

Similar reasoning explains why the estimated mortality rate for pneumonia increases after calibration. The estimate of the InSilicoVA sensitivity for

pneumonia is near 50% which is by far the lowest estimated sensitivity for InSilicoVA estimated by the ensemble model. The EAVA estimated pnuemonia sensitivity is 62% which is similar to the EAVA sensitivites for other infections and other COD, which are the most common GS-COD MITS causes. The estimates of EAVA false positives for pneumonia when the true MITS COD is malaria, other, and other infections, are also higher than the analogous numbers for InSilicoVA. All of this gives insight into why the ensemble pneumonia mortality rate estimate of 31% is much higher than the InSilicoVA estimate (19%) as compared to the EAVA estimate (27%).

Finally, looking at other infections, we see from Figure 5.7 that when the GS-COD is pneumonia, both InSilicoVA and EAVA are expected to predict a 25% probability that we get a false positive for other infections. In addition, when the GS-COD is malaria, EAVA again is expected to predict a 25% probability that the COD is other infections. These misclassification rate of around 25% are higher than the other misclassification rates, which tend to be between 0-12.5%. The fact that both algorithms tend to predict many false positives for other infections gets reflected in both algorithms' uncalibrated estimate (Figure 5.2) where other infections is the leading cause of child death. The results after the ensemble calibration adjusts for the false positives and estimates a lower mortality rate from other infections.

5.4.5 Single-cause-MITS vs multi-cause-MITS-calibration:

As a sensitivity analysis, we also compare the ensemble multi-cause model estimates to the ensemble model single-cause model estimates. Figure 5.8

shows that the CSMF estimates are largely similar between the two methods, with the single-cause MITS data leading to a slightly higher estimated severe malnutrition mortality rate (9% versus 5%), and a slightly lower estimated pneumonia mortality rate (28% versus 31%).

The main driver behind the lower pneumonia mortality rate estimate for the single-cause model is likely to be that the single-cause model estimates higher misclassification to pneumonia from severe malnutrition, while both models estimate similar misclassification rates conditional on the GS-COD being pneumonia. The single-cause model also estimates higher misclassification to other infections from severe malnutrition, especially for InSilicoVA, which is likely the reason for the higher mortality rate estimate for severe malnutrition. Similar reasoning also explains why the single-cause model also estimates a slightly higher mortality rate for HIV.

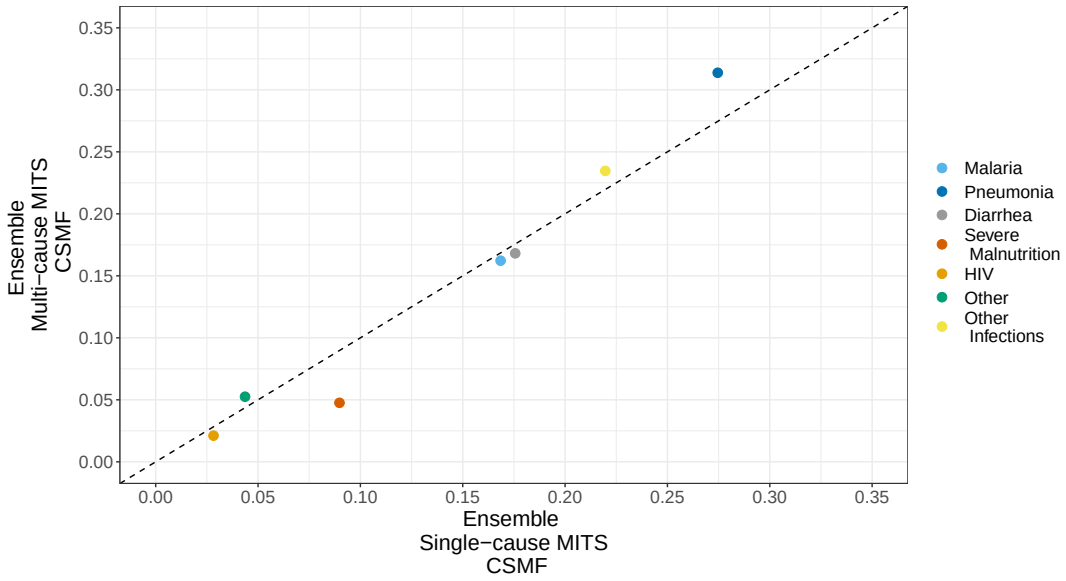


Figure 5.8: Comparison of CSMF estimates from the calibrated ensemble multi-cause MITS model versus the the calibrated ensemble single-cause MITS model.

5.4.6 Model comparison:

Finally, Figure 5.9 compares the WAIC between the calibrated and uncalibrated models, for both the multi-cause and single-cause GS-COD data, where a lower WAIC indicates better model fit. For both the multi-cause and single-cause data, the calibrated models significantly outperform their uncalibrated counterpart. While we are unable to fully evaluate the models on a completely independent validation data set as the calibration does not offer individual predictions but only the calibrated CSMF estimate for the population, WAIC offers a way to do this evaluation using just the collected data, and these results provide strong evidence in favor of using the calibrated models for estimating the CSMF, as opposed to using the uncalibrated CSMF estimate.

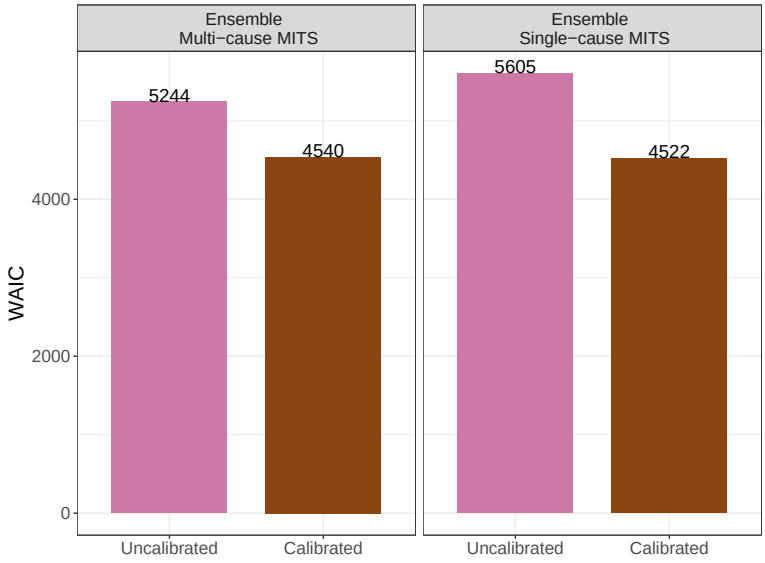


Figure 5.9: WAIC for the calibrated and uncalibrated ensemble models, using both multi-cause and single-cause MITS data.

5.5 Discussion

This study outlines a complete pipeline to use a limited dataset of paired VA records and GS-COD (in this case from MITS) to calibrate CSMF estimates obtained from VA algorithms applied to abundant unpaired VA data from a nationally representative survey. The calibration works with both multiple and single-cause outputs from both the VA and the MITS. We show that for child deaths in Mozambique, this results in higher estimated mortality from pneumonia and malaria, and lower estimated mortality from other infections. We also show that this calibrated model outperforms the uncalibrated model in terms of WAIC.

We also give insight into why the calibration model resulted in these changes to the CSMF for this application. However, giving a simple explanation for the changes made to the CSMF may not always be possible as the calibration is reflect the total change affected by many different misclassification rates. This underscores the need for clear communication between statistical practitioners and government officials and stakeholders to understand the general principles of the calibration model which are intuitive and interpretable.

As both the COMSA and CHAMPS are ongoing surveillance projects, we expect the size of both data sources to increase over time. While we expect that this will result in more precise CSMF estimates, it presents an interesting challenge in how to interpret the changing estimates. With a larger CHAMPS dataset, we expect the prior to play less of a role in estimating the misclassification rates. Given the uncalibrated misclassification rate estimates in Figure 5.3

that show extremely low sensitivities, we would expect the CSMF estimates to change as the posterior means for the misclassification rates become closer to the uncalibrated estimates. However, the true mortality rates may also be changing as a function of health interventions within Mozambique. Thus, an important future direction is to develop methods to determine what sources are influencing the changing CSMF estimates.

Another important future direction for research is to develop methods for comparing calibrated model estimates, for example comparing the estimate from the ensemble calibration model to just the calibrated InSilicoVA model. The difficulty in this is that there is no external data for model validation, and that each calibration model using a different VA algorithm uses a different source of VA data. However, we have shown that calibrated models outperform their uncalibrated counterparts in terms of WAIC for the COMSA and CHAMPS data. In addition, extensive simulation studies have shown the superiority of the ensemble model to models which only use one VA algorithm (Datta et al., 2020; Fiksel et al., 2020).

Despite these important unsolved challenges for producing calibrated CSMF estimates, we believe that this method produces more informed CSMF estimates than simply aggregating VA algorithm predictions especially given the large misclassification rates we observe for both VA algorithms. As more countries begin implementing VAs within national surveillance systems, they should also invest in obtaining a smaller number of deaths with both VA and GS-COD information, even if there is uncertainty in the GS-COD. Projects such as the global symptom-cause archive (Clark, Setel, and Li, 2019) may

help to establish misclassification rates for many algorithms and regions of the world in order to produce accurate COD information for low and middle income countries.

5.6 Supplemental Figures

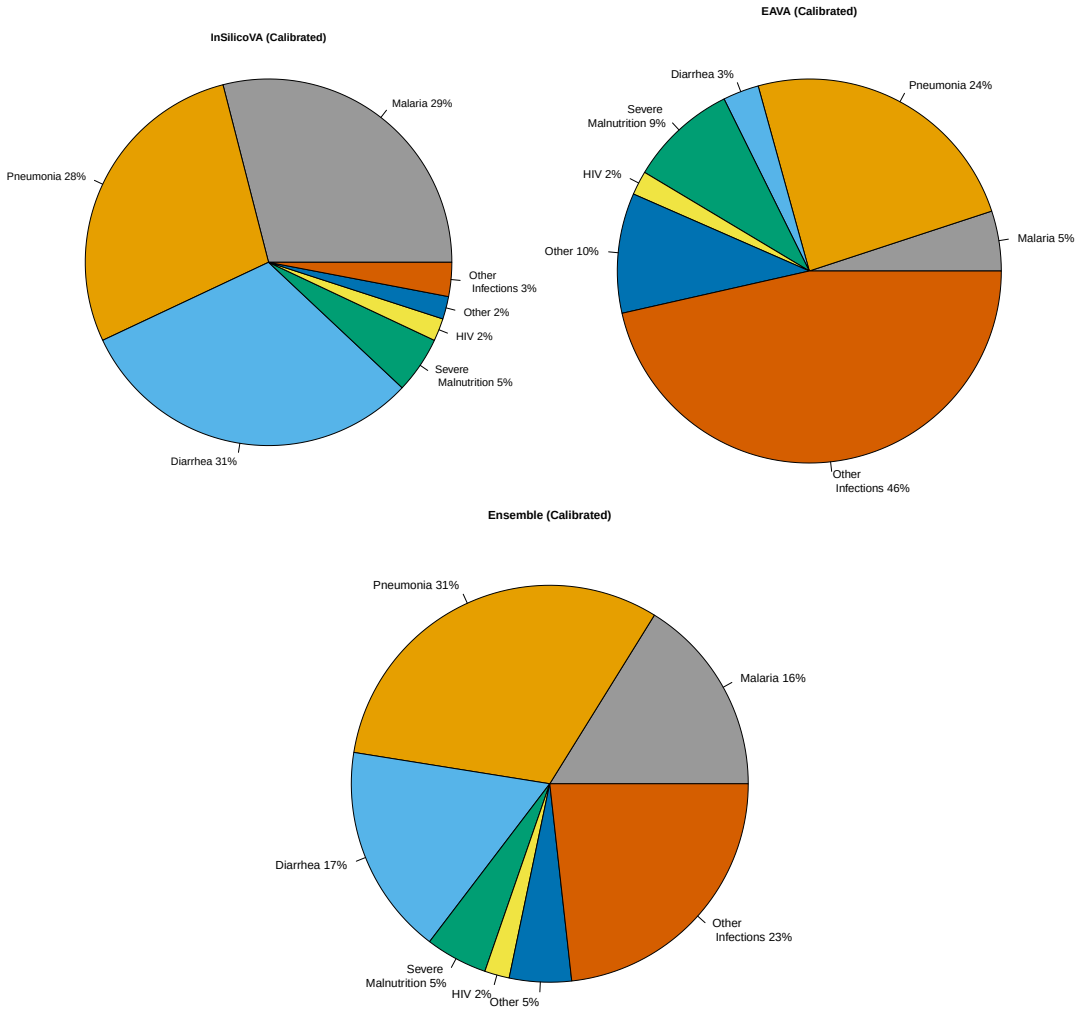


Figure 5.10: Multi-cause calibrated CSMFs for InSilicoVA, EAVA, and the ensemble method.

References

Nichols, Erin K, Peter Byass, Daniel Chandramohan, Samuel J Clark, Abraham D Flaxman, Robert Jakob, Jordana Leitao, Nicolas Maire, Chalapati Rao, Ian Riley, et al. (2018). "The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0". In: *PLoS medicine* 15.1.

Soleman, Nadia, Daniel Chandramohan, and Kenji Shibuya (2006). "Verbal autopsy: current practices and challenges". In: *Bulletin of the World Health Organization* 84, pp. 239–245.

McCormick, Tyler H, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark (2016). "Probabilistic cause-of-death assignment using verbal autopsies". In: *Journal of the American Statistical Association* 111.515, pp. 1036–1049.

Byass, Peter, Daniel Chandramohan, Samuel J Clark, Lucia D'ambruoso, Edward Fottrell, Wendy J Graham, Abraham J Herbst, Abraham Hodgson, Sennen Hounton, Kathleen Kahn, et al. (2012). "Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool". In: *Global health action* 5.1, p. 19281.

Miasnikof, Pierre, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha (2015). "Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths". In: *BMC medicine* 13.1, p. 286.

Kalter, Henry D, Jamie Perin, and Robert E Black (2016). "Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death". In: *Journal of global health* 6.1.

Jha, Prabhat, Dinesh Kumar, Rajesh Dikshit, Atul Budukh, Rehana Begum, Prabha Sati, Patrycja Kolpak, Richard Wen, Shyamsundar J Raithatha, Utkarsh Shah, et al. (2019). "Automated versus physician assignment of

cause of death for verbal autopsies: randomized trial of 9374 deaths in 117 villages in India". In: *BMC medicine* 17.1, p. 116.

Clark, Samuel J, Zehang Li, and Tyler H McCormick (2018). "Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for Verbal Autopsy". In: *arXiv preprint arXiv:1803.07141*.

Datta, Abhirup, Jacob Fiksel, Agbessi Amouzou, and Scott L Zeger (2020). "Regularized Bayesian transfer learning for population-level etiological distributions". In: *Biostatistics*. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxaa001. eprint: https://academic.oup.com/biostatistics/advance-article-pdf/doi/10.1093/biostatistics/kxaa001/32412903/kxaa001.pdf. URL: https://doi.org/10.1093/biostatistics/kxaa001.

Fiksel, Jacob, Abhirup Datta, Agbessi Amouzou, and Scott Zeger (2020). "Generalized Bayesian Quantification Learning". In: *arXiv preprint arXiv:2001.05360*.

Byass, Peter (2016). "Minimally invasive autopsy: a new paradigm for understanding global health?" In: *PLoS medicine* 13.11.

CHAMPS Cause of Death Data. https://champshealth.org/cause-of-death-data-visualization/.

Castillo, Paola, Miguel J Martínez, Esperança Ussene, Dercio Jordao, Lucilia Lovane, Mamudo R Ismail, Carla Carrilho, Cesaltina Lorenzoni, Fabiola Fernandes, Rosa Bene, et al. (2016). "Validity of a minimally invasive autopsy for cause of death determination in adults in Mozambique: an observational study". In: *PLoS medicine* 13.11.

Watanabe, Sumio (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory". In: *Journal of Machine Learning Research* 11.Dec, pp. 3571–3594.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and computing* 27.5, pp. 1413–1432.

Clark, Samuel J, Philip Setel, and Zehang Li (2019). "Verbal Autopsy in Civil Registration and Vital Statistics: The Symptom-Cause Information Archive". In: *arXiv preprint arXiv:1910.00405*.

Chapter 6

Discussion and Conclusion

This thesis has developed important methodological advancements that allow for improved CSMF estimates from VAs. The GBQL approach presented in Chapter 4 is especially important in advancing the area of quantification and mortality estimation, as it provides a robust Bayesian framework to incorporate compositional predictions and labels. Our Gibbs sampler using rounding and coarsening allows for fast posterior sampling, which encourages the use of our method by policymakers and stakeholders. This Gibbs sampler is implemented in the 'CalibratedVA' R-package. We have shown how the GBQL approach can be used for CSMF estimation in Mozambique. We believe this approach can and should be adopted by future demographic surveillance systems that rely on VAs for mortality data.

While the GBQL approach uses the loss function presented in Chapter 3 for quantification with compositional data, we should how this loss function can be used for transformation-free regression with compositional data. The direct regression model provides simple coefficient interpretation, as opposed to previous compositional regression models, and is easy to implement. In

Chapter 3, we applied this model to research data from education and medical imaging, and in Chapter 5 we applied this model to estimate the uncalibrated misclassification rates in the multi-cause-VA-multi-cause-MITS model.

This work has opened several future areas of research. While the direct regression model in Chapter 3 is an important contribution to the field of compositional data analysis, it does not currently allow for inclusion of confounding variables. Ideally, the model would be able to handle confounding variables of mixed data types, rather than just compositional confounding variables. There are two difficulties in including additional variables into the model. First, because our model is represented as a single Markov transition, it allows for simple coefficient interpretation. If we were to include confounding variables in the model, the coefficients should interpret in a similar manner, although it would be conditional on the values of the confounding variables. Second, any model with compositional data must respect the unit-sum restriction of this data. In the direct regression model, we perform parameter estimation with constrained optimization, and the compositional data allowed us to use a simple EM algorithm. However, by including confounding variables, the constrained optimization problem becomes more complex in order to ensure that the expected value of the outcome is compositional for each combination of covariates. The fact that one or more of the covariates will also be compositional, and thus linearly dependent, further complicates the computation.

The use of Bayesian updating with loss-functions for compositional data in Chapter 4 also opens the door for generalized Bayesian regression for

compositional outcomes. This could be a Bayesian extension of the direct regression model in Chapter 3, or incorporating priors into the MQL regression model with continuous covariates.

While we compared the calibrated and uncalibrated models using the WAIC in Chapter 5, we do not currently have a method to compare calibrated models using different combinations of CCVAs. The reason for this is that the WAIC is an estimate of the log-likelihood for *future data*. Comparing a calibrated ensemble model with a calibrated model using just InSilicoVA predictions is difficult, as the ensemble model must account for both InSilicoVA and EAVA predictions, but the latter model only has to account for InSilicoVA predictions. Although our simulations in Chapters 2 and 4 show that the ensemble model performs well for CSMF estimation from the PHMRC data, it is important to develop metrics that allow for model selection in order to communicate to stakeholders which CSMF estimate should be used. Furthermore, improved metrics for comparing calibrated models would allow for more principled selection of the prior parameters.

Finally, Chapter 5 showed that labeled data may come from multiple countries, and not just the country of interest. An extension of the GBQL model could be to add an additional hierarchical structure to estimation of the misclassification matrix, that allows for country-specific misclassifications. One could decide how much information is pooled between countries to estimate an "overall" misclassification matrix that is used for calibration.

Jacob Fiksel

Contact Information	1019 N. Calvert St Baltimore, MD 21202	608-345-8988 jfiksel@gmail.com

Research Interests

Applied Bayesian statistical modeling, global public health, cancer genomics

Education

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Ph.D., Biostatistics, *Expected:* Spring 2020

- Thesis Topic: Bayesian Quantification Learning with Applications to Mortality Surveillance
- Advisor: Abhirup Datta Ph.D

Pomona College, Claremont, CA

B.S., Mathematics with a Statistics focus, May 2015

- *Cum Laude*
- GPA 3.85/4.00
- Distinction in the Senior Exercise

Research Experience

Johns Hopkins Bloomberg School of Public Health

- Developed a novel Bayesian framework using generalized Gibbs updates for robust quantification
- As part of the Countrywide Mortality Surveillance for Action (COMSA) in Mozambique team, built the calibratedVA R package for application of the Bayesian quantification framework to estimation of the leading causes of neonatal and child death in Mozambique.
- Assisted in the development and implementation of statistical methods to detect chromosomal abnormalities in circulating cell-free DNA collected from cancer patients. This work resulted in a Nature publication
- In collaboration with a clinical radiation oncologist, constructed a Random Survival Forest that uses clinical data from patients with bone metastases to make personalized estimates of survival, allowing for more informed decision making in the choice of palliative radiation treatment regimen
- Developed widely-used teaching guides for the use of GitHub Classroom in teaching undergraduate statistics.

Refereed Journal Publications

1. Jillian Phallen, Alessandro Leal, Brian D Woodward, Patrick M Forde, Jarushka Naidoo, Kristen A Marrone, Julie R Brahmer, **Jacob Fiksel**, Jamie E Medina, Stephen Cristiano, et al. Early noninvasive detection of response to targeted therapy in non–small cell lung cancer. *Cancer research*, 79(6):1204–1213, 2019

2. Stephen Cristiano*, Alessandro Leal*, Jillian Phallen*, **Jacob Fiksel***, Vilmos Adleff, Daniel C Bruhm, Sarah Østrup Jensen, Jamie E Medina, Carolyn Hruban, James R White, et al. Genome-wide cell-free dna fragmentation in patients with cancer. *Nature*, page 1, 2019

3. **Jacob Fiksel**, Leah R Jager, Johannna S Hardin, and Margaret A Taub. Using github classroom to teach statistics. *Journal of Statistics Education*, pages 1–10, 2019

4. Sara R Alcorn, Christen Elledge, Jean L Wright, Thomas J Smith, Todd R McNutt, **Jacob Fiksel**, Scott Zeger, and Theodore L DeWeese. Frequency of complicated symptomatic bone metastasis over a breadth of operational definitions. *International Journal of Radiation Oncology · Biology · Physics*, 2019

* Co-first author

5. Abhirup Datta, **Jacob Fiksel**, Agbessi Amouzou, and Scott Zeger. Regularized Bayesian transfer learning for population level etiological distributions. *arXiv e-prints*, page arXiv:1810.10572, Oct 2018. Accepted in *Biostatistics*

Papers in Review

1. **Jacob Fiksel**, Abhirup Datta, Agbessi Amouzou, and Scott Zeger. Generalized Bayesian Quantification Learning. *arXiv e-prints*, page arXiv:2001.05360, Jan 2020
2. Sara R. Alcorn, **Jacob Fiksel**, Jean L. Wright, Christen R. Elledge, Thomas J. Smith, Powell Perng, Sarah Saleemi, Todd McNutt, Theodore L. DeWeese, and Scott L. Zeger. Developing an Improved Statistical Approach for Survival Estimation in Bone Metastases Management: The Bone Metastases Ensemble Trees for Survival (BMETS) Model. Under review at *International Journal of Radiation Oncology · Biology · Physics*

Papers in Preparation

1. **Jacob Fiksel**, Abhirup Datta, and Scott Zeger. A Transformation-free Linear Regression for Compositional Outcomes and Predictors

Computing knowledge

- R coding and package development
- Stan, JAGS, and custom MCMC samplers
- Development of machine learning models
- Next generation sequencing analysis with command line tools and R
- Reproducible computing with Git, GitHub, and GitHub Classroom
- Simulation studies using a high performance computing exchange (JHPCE)

Awards

JHSPH Department of Biostatistics Helen Abbey Award, 2020

- For excellence in teaching

Presentations

Johns Hopkins Research on Aging Showcase Poster Competition April 2018

- *Optimized Survival Evaluation to Guide Bone Metastases Management: Developing an Improved Statistical Approach.* **1st place in the graduate student poster competition**

Teaching Experience

Lead Teaching Assistant

- Fall 2018, 2019: Statistical Methods in Public Health I-II
  - Lead weekly labs 2-3 times per week for 10-40 graduate students

Teaching Assistant

- Spring 2018: Analysis of Longitudinal Data and Multilevel Statistical Models in Public Health
- Spring 2017: Statistical Methods in Public Health III-IV
- Fall 2016, 2018: Biostatistics for Public Health
  - Lead weekly labs once per week for 25 undergraduate students

Service

JHSPH Biostatistics Computing Club Co-President September 2016-June 2017

- Organized bi-weekly student presentations on computing related topics for the Biostatistics department

Recruiting Committee, Division of Biostatistics December 2016 – Present

- Assist with planning of annual Division of Biostatistics Open House and Admitted Student Visit Days
- Meet with prospective and admitted students

Abhirup Datta
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
E-mail: abhidatta@jhu.edu

Scott Zeger
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
E-mail: sz@jhu.edu

Robert Scharpf
Associate Professor
Department of Oncology
Johns Hopkins University
E-mail: rscharpf@jhu.edu